Journal of Econometrics xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

High dimensional regression coefficient test with high frequency data

Dachuan Chen [a], Long Feng [b,*], Per A. Mykland [c], Lan Zhang [d]

[a] School of Economics, Singapore Management University, Singapore
[b] School of Statistics and Data Science, KLMDASR, LEBPS, and LPMC, Nankai University, China
[c] Department of Statistics, The University of Chicago, United States of America
[d] Department of Finance, University of Illinois at Chicago, United States of America

ARTICLE INFO

JEL classification:
C12

Keywords:
High dimensionality
Time-varying regression coefficient process
High frequency data
Hypothesis tests
Sum-type test
Max-type test
Asymptotic independence
Fisher's combination test

ABSTRACT

This paper presents the first study on high-dimensional regression coefficient tests with high-frequency financial data. These tests allow the number of regressors to be larger than the number of observations within each estimation block and can grow to infinity in asymptotics. In this paper, the sum-type test and max-type test have been proposed, where the former is suitable for the dense alternative (many small betas) and the latter is suitable for the sparse alternative (a very small number of large betas). By showing the asymptotic independence between the sum-type test and max-type test, the paper proposes a third test – Fisher's combination test, which is robust to both dense and sparse alternatives. The paper derives the limiting null distributions of the three proposed tests and analyzes the asymptotic behavior of their powers. Monte Carlo simulations demonstrate the validity of the theoretical results developed in this paper. Empirical study shows the impact of high frequency (HF) factors when being added to a Fama–French-style factor model. We found that the HF effects are time varying. The proposed tests can help identify those time periods when the HF factors carry (significant) incremental information for the test asset. Our tests could shed light on market timing in a trading strategy.

1. Introduction

High frequency regression (or, realized regression) has been receiving increasing attention in recent years. Enabled by the continuous-time modeling approach, the nonparametric framework of high frequency regression allows for time-varying features for the covariance structure of the dependent variable process and the independent variable processes, and permits the time-varying regression coefficients (or, betas). The analysis of high frequency regression has wide applications in finance, for example, (i) it can be helpful in characterizing the error of the hedging strategy in financial trading (e.g., see Mykland and Zhang (2006)); (ii) it can be applied to asset pricing based on the time-varying (or, conditional) factor models (e.g., see Aït-Sahalia et al. (2020)); (iii) it is also helpful in risk management and portfolio allocation through the accurate estimation of large covariance and precision matrices (e.g., see Fan et al. (2016) and Chen et al. (2024)).

The statistical methodologies developed for high frequency regression mainly focus on the following areas: (i) the estimation of regression coefficients (betas), including the realized beta estimator with single regressor (e.g., see Barndorff-Nielsen and Shephard (2004) and Andersen et al. (2005)), the integrated beta estimator with multiple regressors (e.g., see Mykland and Zhang (2009), Aït-Sahalia et al. (2020) and Chen et al. (2024)); (ii) the estimation of idiosyncratic volatilities, e.g., see Mykland and Zhang (2006)

[*] Corresponding author.
E-mail address: flnankai@nankai.edu.cn (L. Feng).

https://doi.org/10.1016/j.jeconom.2024.105812
Received 19 January 2023; Received in revised form 13 November 2023; Accepted 16 May 2024

and Aït-Sahalia et al. (2020); (iii) the estimation of large covariance and precision matrices, e.g., see Fan et al. (2016) and Chen et al. (2024); and (iv) the testing problem about the time-variation in the beta processes, e.g., see Reißet al. (2015) and Kong and Liu (2018).

All aforementioned theories are developed based on the assumption that the number of regressors is finite in the asymptotic setting. However, in the era of big data and machine learning, the number of regressors becomes larger and larger (for example, there are more than 200 high frequency factors available currently). In this high dimensional setting, the covariance matrix estimator of the regressors is very likely to be not invertible. Thus, it is necessary to develop the statistical inference methods for the regression coefficient processes with high dimensionality. A frequently encountered challenge in high dimension regression is the detection of relevant variables. Recently, Kim and Shin (2022) proposed a Thresholding dEbaised Dantzig (ETD) estimator for the high dimensional high frequency factor models, which can simultaneously select and estimate the regression coefficient process. Shin and Kim (2023) proposed the robust estimation procedure to explain the heavy-tailed observations. Hypothesis testing is another effective method for identifying relevant variables. As far as we know, this paper is the first one to conduct the high dimensional regression coefficient tests with high frequency data.

The theoretical contribution of this paper is three-fold. First, the existing high dimensional hypothesis tests under the low frequency setting are mainly designed for the independent and identically distributed observations or the stationary observations. The typical problem in the low frequency setting is to test the high dimensional mean vector, and the existing methodologies can be classified into three types: (i) max-type test, see e.g., Cai et al. (2014); (ii) sum-type test, see e.g., Bai and Saranadasa (1996), Srivastava and Du (2008), Srivastava (2009) , Chen and Qin (2010) and Srivastava et al. (2013); (iii) combination test, see, e.g., Fan et al. (2015), Xu et al. (2016), He et al. (2021), Feng et al. (2022) and Chen and Feng (2022). In contrast, the continuous time modeling framework enables us to relax these assumptions, for example, to assume that the covariance matrix of the independent variable processes and the variance of the residual process are time-varying. Second, under the low frequency setting, the high dimensional beta test mainly addresses a much simpler and known problem where the beta vector is deterministic and constant, see, e.g., Liu et al. (2020). In contrast, the regression coefficients in our paper are time-varying and stochastic. Third, the traditional statistical techniques and methodologies under the low frequency setting cannot be directly applied to the inference related to continuous time models, i.e., U-statistics, likelihood method, etc. Therefore, novel methodologies are developed in this paper to conduct the high dimensional beta tests in the high frequency setting.

Under the framework of high frequency regression, we propose tests for the high dimensional regression coefficient that account for the time-varying feature in both the regression coefficient processes and the covariance structure of the regressors. In this setting, the high dimensionality implies that the number of regressors can exceed the number of observations in each estimation block and can also diverge to infinity asymptotically. In this paper, we have developed the sum-type test and max-type test for high dimensional regression coefficient processes. Furthermore, the asymptotic independence between the sum-type test and the max-type test is established, leading to the Fisher's combination test. In the theoretical development, we derive the asymptotic distributions of these three test statistics under the null hypothesis, and analyze the asymptotic behavior of their powers. Monte Carlo simulations demonstrate the validity of the theoretical results of the three tests. We also perform an empirical study based on the intraday prices of the components of S&P 100 Index and the prices of a few Exchange-Traded Funds (ETF) from Jan 2007 to Dec 2017. We construct three scenarios of the realized regressions, which serve as the dense alternative, the sparse alternative, and the null hypothesis, respectively. The result shows that (i) our Fisher's combination test is robust with respect to both dense and sparse alternatives; (ii) all of the three proposed tests are non-significant under the null hypothesis. In another application, we investigate the role of high frequency (HF) factors when being added to a traditional Fama–French (FF) factor model. We found that from a long horizon (say, 11 years in our data), HF factors add highly significant contribution to explaining the test assets (various ETFs). When viewed from a monthly level, the HF factors becomes much weaker, especially in a Fama–French-style 6-factor model. We also found that impact of HF factors are time-varying. During certain time periods, HF factors could carry some incremental information when included in the Fama–French 3 factor model. For example, in the first quarter of 2012, our tests detected many low-impact HF factors influencing the returns of the health care ETF XLV.

The theoretical results in the current paper touch upon both high frequency and high dimensional data analysis. Our work relates to high frequency principal component analysis (PCA) and high frequency regression, which can be applied into the estimation of high dimensional covariance and precision matrices, e.g., see Aït-Sahalia and Xiu (2017), Pelger (2019, 2020), Kong (2017), Bollerslev et al. (2019) , Dai et al. (2019), Chen et al. (2020) and Kong et al. (2021) for high frequency PCA; and Fan et al. (2016), Dai et al. (2019) and Chen et al. (2024) for high frequency regression. Another relevant literature is the test for constant factor loading matrix, where the number of dependent variables can grow to infinity, e.g., see Kong and Liu (2018).

This paper is organized as follows. Section 2 provides the basic settings about the model. Section 3 sets up the null hypothesis and the alternative hypothesis, and proposes the three types of the high dimensional regression coefficient tests, including the sum-type test, max-type test and Fisher's combination test. Section 4 shows the Monte Carlo evidence of the proposed theoretical methodology. Section 5 provides the empirical study based on the proposed tests. Section 6 concludes this paper. All mathematical proofs of the theoretical results in this paper are collected in the online supplementary material.

We introduce several notations as follows. $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ denote the smallest and the largest eigenvalues of the matrix $\mathbf{A}$, respectively. For a $p \times q$ matrix $\mathbf{A}$, define $\|\mathbf{A}\|_{\max} = \max_{1\le j\le p, 1\le k\le q} \left|\mathbf{A}^{(j,k)}\right|$. For random vectors $\mathbf{X}$ and $\mathbf{Y}$, $\bar{S}_1(\mathbf{X})$ denotes the first order Stein discrepancy as defined in Definition 1.3 of Fathi (2021), and $W_2(\mathbf{X}, \mathbf{Y})$ denotes the Wasserstein distance between $\mathbf{X}$ and $\mathbf{Y}$ as defined in Section 3 of Ledoux et al. (2015). For a random variable X, $\|X\|_2 = \left[E\left(X^2\right)\right]^{1/2}$. For a $l \times 1$ vector $\mathbf{a} = \left(a_1, a_2, \dots, a_l\right)^\intercal$, $\|\mathbf{a}\|$ denotes the Euclidean norm of $\mathbf{a}$, i.e., $\|\mathbf{a}\| = \left(\sum_{k=1}^{l} a_k^2\right)^{1/2}$. $N_d(\mu, \Sigma)$ denotes the d-dimensional normal random vector with mean vector μ and covariance matrix Σ and $\mathbb{I}_d$ is the $d \times d$ identity matrix.

2. Models

We assume that the $(p+1)$-dimensional process $\Xi_t = \left(X_t^{(1)}, X_t^{(2)}, \dots, X_t^{(p)}, Y_t\right)^\intercal$ is a continuous Itô -semimartingale as follows:

$$d\Xi_t = \mu_t dt + \sigma_t dW_t,$$

where μ_t is a $(p+1)\times 1$ vector process and σ_t is a symmetric and invertible $(p+1)\times(p+1)$ matrix process, and μ_t and σ_t are Itô processes, W_t is a $(p+1)$-dimensional standard Brownian motion. Here we introduce a notation X_t as $X_t = \left(X_t^{(1)}, X_t^{(2)}, \dots, X_t^{(p)}\right)^\intercal$.

The spot covariance process of Ξ_t at time t can be expressed as follows:

$$c_t = \left(\sigma\sigma^\intercal\right)_t, \tag{2.1}$$

which belongs to the set of positive-semidefinite matrices for any $0 \le t \le \mathcal{T}$.

Note that the spot covariance process can be further partitioned as follows:

$$c_t = \begin{pmatrix} c_t^{X,X} & c_t^{Y,X} \\ c_t^{X,Y} & c_t^{Y,Y} \end{pmatrix},$$

where the $p\times p$ spot covariance matrix process of X_t is denoted by $c_t^{X,X}$, and $c_t^{Y,X} = \left(c_t^{X,Y}\right)^\intercal$.

Assume that the processes $Y_t, X_t^{(1)}, \dots, X_t^{(p)}$ are observed on an equidistant grid: $\left\{t_{n,j}\right\}_{0\le j\le n}$ with $t_{n,j} = j\Delta t_n$ and $\Delta t_n = \mathcal{T}/n$. We define $\Delta X_{t_{n,j+1}} = X_{t_{n,j+1}} - X_{t_{n,j}}$. Similarly, we can define $\Delta Y_{t_{n,j+1}}, \Delta Z_{t_{n,j+1}}$ and $\Delta t_{n,j+1}$.

We also introduce a coarser grid $\left\{\tau_{n,i}\right\}_{0\le i\le K_n}$ where $\tau_{n,i} = iM_n\Delta t_n$ and M_n is a positive integer such that $M_n = n/K_n$. In asymptotics, we assume $M_n = O(1)$ and $K_n = O(n)$ as $n \to \infty$.

3. High-dimensional regression coefficient tests

Suppose that the processes are related by

$$dY_t = \sum_{k=1}^{p} \beta_t^{(k)} dX_t^{(k)} + dZ_t \text{ with } \left\langle X^{(k)}, Z\right\rangle_t = 0 \text{ for all } 0 \le t \le \mathcal{T} \text{ and } 1 \le k \le p. \tag{3.1}$$

The symbol $\langle\cdot,\cdot\rangle_t$ denotes the quadratic covariation of two Itô processes. Here we provide an assumption for the residual process Z_t.

Assumption 1. Suppose the stochastic process Y_t follows the continuous-time multiple regression model as expressed by (3.1), in which the factor loading processes $\beta_t^{(k)}, 1 \le k \le p$ are Itô processes. Assume the residual process Z_t satisfying

$$dZ_t = \mu_t^Z dt + \sigma_t^Z dW_t^Z,$$

and assume the drift process μ_t^Z and the volatility process σ_t^Z are also Itô processes. □

Based on above assumption, it is easy to see that $\langle Z, Z\rangle_t = \int_0^t \left(\sigma_s^Z\right)^2 ds$, and $\langle Z, Z\rangle_t' = \left(\sigma_t^Z\right)^2$.

Remark 1 (*Discussion About Possible Latent Factors*)**.** In the setting of regression, we only consider the factors that are observable. However, there might be latent factors that are not included into the regression equation. We discuss the latent factors in the following two cases. (i) If the latent factor F_t is orthogonal to the independent variables X_t, i.e., $\langle F, X\rangle_t' \equiv 0$ for $0 \le t \le \mathcal{T}$, this latent factor can be absorbed into the residual process Z_t. (ii) If the latent factor F_t is correlated with the independent variables X_t, i.e., $\langle F, X\rangle_t' \ne 0$, X_t can be driven by a set of common latent factors and these common latent factors can be pervasive (Fan et al., 2013). This case has wide applications in the financial market, for example, X_t is the asset returns of a large number of individual stocks, which are actually driven by a small number of common factors, see, e.g., Fama and French (2017) for the low frequency version and Aït-Sahalia et al. (2020) for the high frequency version. The proposed tests in this paper can also cover this case, as explained in Remark 3. □

To develop our theory, we partition the vector X_t of p independent variables into two subsets: group a and group b. Group a contains the first q variables, while group b contains the remaining $p-q$ variables. Thus, we can introduce the subsequent notations: $X_t = (X_{a,t}^\intercal, X_{b,t}^\intercal)^\intercal$ where $X_{a,t} = \left(X_t^{(1)}, \dots, X_t^{(q)}\right)^\intercal$ and $X_{b,t} = \left(X_t^{(q+1)}, \dots, X_t^{(p)}\right)^\intercal$. Here, following the definition of $\Delta X_{t_{n,j+1}}$, we define $\Delta X_{a,t_{n,j+1}} = X_{a,t_{n,j+1}} - X_{a,t_{n,j}}$ and $\Delta X_{b,t_{n,j+1}} = X_{b,t_{n,j+1}} - X_{b,t_{n,j}}$. Further, we could also introduce a notation for the vector of regression coefficients: $\boldsymbol{\beta}_t = (\boldsymbol{\beta}_{a,t}^\intercal, \boldsymbol{\beta}_{b,t}^\intercal)^\intercal$ where $\boldsymbol{\beta}_{a,t} = \left(\beta_t^{(1)}, \dots, \beta_t^{(q)}\right)^\intercal$ is a $q\times 1$ column vector process denoting the regression coefficients of group a variables, and $\boldsymbol{\beta}_{b,t} = \left(\beta_t^{(q+1)}, \dots, \beta_t^{(p)}\right)^\intercal$ is a $(p-q)\times 1$ column vector process denoting the regression coefficients of group b variables. For group a, we assume the number of variables q to be fixed in asymptotics and we assume that the group a variables are always included in the realized regression, i.e., $\boldsymbol{\beta}_{a,t} \ne 0$ for some $0 \le t \le \mathcal{T}$. For group b, we assume the number of variables $p-q$ to be diverging and we are interested in testing whether $\boldsymbol{\beta}_{b,t} \equiv 0$. In the high-dimensional scenario, we assume that $q < M_n$ and p is much larger than M_n.

Here the $\beta_t^{(k)}$ is the factor loading with respect to the covariate process $X_t^{(k)}$. If $\beta_t^{(k)} = 0$, there is no relationship between the dependent process Y_t and the covariate process $X_t^{(k)}$. In order to test whether those covariates in $X_{b,t}$ are related with Y_t, we consider the following global testing problem:

$$H_0 : \boldsymbol{\beta}_{b,t} \equiv \mathbf{0} \text{ for all } t \in \left[0, \mathcal{T}\right] \text{ vs. } H_1 : \exists t \in \left[0, \mathcal{T}\right] \text{ such that } \boldsymbol{\beta}_{b,t} \neq \mathbf{0}. \tag{3.2}$$

If we do not reject the null hypothesis, we can significantly reduce the number of covariate process in model (3.1), thereby improving the accuracy of statistical inference.

Define

$$\mathcal{Y}_i = \begin{pmatrix} \Delta Y_{t_{n,(i-1)M_n+1}} \\ \Delta Y_{t_{n,(i-1)M_n+2}} \\ \vdots \\ \Delta Y_{t_{n,iM_n}} \end{pmatrix}_{M_n \times 1}, \mathcal{X}_i = \begin{pmatrix} \Delta X^{\intercal}_{t_{n,(i-1)M_n+1}} \\ \Delta X^{\intercal}_{t_{n,(i-1)M_n+2}} \\ \vdots \\ \Delta X^{\intercal}_{t_{n,iM_n}} \end{pmatrix}_{M_n \times p} \text{ and } \mathcal{Z}_i = \begin{pmatrix} \Delta Z_{t_{n,(i-1)M_n+1}} \\ \Delta Z_{t_{n,(i-1)M_n+2}} \\ \vdots \\ \Delta Z_{t_{n,iM_n}} \end{pmatrix}_{M_n \times 1}. \tag{3.3}$$

In the low-dimensional setting, say $M_n > p$, the least square estimator for the spot beta (see Mykland and Zhang (2009) and Aït-Sahalia et al. (2020)),

$$\hat{\boldsymbol{\beta}}_{\tau_{n,i-1}} = \left(\mathcal{X}_i^{\intercal}\mathcal{X}_i\right)^{-1}\mathcal{X}_i^{\intercal}\mathcal{Y}_i.$$

However, the estimator $\hat{\boldsymbol{\beta}}_{\tau_{n,i-1}}$ and related central limit theorem cannot work when $M_n < p$ because $\mathcal{X}_i^{\intercal}\mathcal{X}_i$ is not invertible.

In this paper, we will develop the high dimensional regression coefficient test where p can be much larger than M_n, and p can also grow to infinity in asymptotics. We further define:

$$\mathcal{X}_{a,i} = \left(\Delta X_{a,t_{n,(i-1)M_n+1}}, \Delta X_{a,t_{n,(i-1)M_n+2}}, \dots, \Delta X_{a,t_{n,iM_n}}\right)^{\intercal} \tag{3.4}$$

and

$$\mathcal{X}_{(j),i} = \left(\Delta X^{(j)}_{t_{n,(i-1)M_n+1}}, \Delta X^{(j)}_{t_{n,(i-1)M_n+2}}, \dots, \Delta X^{(j)}_{t_{n,iM_n}}\right)^{\intercal} \tag{3.5}$$

where $\mathcal{X}_{a,i}$ is a $M_n \times q$ matrix and $\mathcal{X}_{(j),i}$ is a $M_n \times 1$ vector.

Thus, we construct a test statistic as follows:

$$V_{j,i} = \Delta t_n^{-1}\mathcal{Y}_i^{\intercal}\tilde{\mathcal{X}}_{(j),i}\left(\tilde{\mathcal{X}}_{(j),i}^{\intercal}\tilde{\mathcal{X}}_{(j),i}\right)^{-1}\tilde{\mathcal{X}}_{(j),i}^{\intercal}\mathcal{Y}_i \text{ for } q+1 \le j \le p,$$

where

$$\mathcal{H}_{a,i} = \mathcal{X}_{a,i}\left(\mathcal{X}_{a,i}^{\intercal}\mathcal{X}_{a,i}\right)^{-1}\mathcal{X}_{a,i}^{\intercal} \text{ and } \tilde{\mathcal{X}}_{(j),i} = \left(\mathbb{I}_{M_n} - \mathcal{H}_{a,i}\right)\mathcal{X}_{(j),i}.$$

To aggregate the test statistics $V_{j,i}$ across blocks, we define:

$$V_j = \frac{1}{K_n}\sum_{i=1}^{K_n} V_{j,i} \text{ for } q+1 \le j \le p.$$

Remark 2 (*Main Idea Concerning Test Statistic Construction*)**.** Based on the definition $\tilde{\mathcal{X}}_{(j),i} = \left(\mathbb{I}_{M_n} - \mathcal{H}_{a,i}\right)\mathcal{X}_{(j),i}$ and the fact $\mathcal{X}_{a,i}^{\intercal}\tilde{\mathcal{X}}_{(j),i} = 0$, $\tilde{\mathcal{X}}_{(j),i}$ is the component of $\mathcal{X}_{(j),i}$ which is orthogonal to the group a variables $\mathcal{X}_{a,i}$, from a geometric perspective. As illustrated in the definition of $V_{j,i}$, it is the sum of squared fitted values (SSFV) of the following regression problem: $\mathcal{Y}_i = \tilde{\mathcal{X}}_{(j),i}\tilde{\beta}_{(j),i} + \tilde{\mathcal{Z}}_{(j),i}$. When $\beta_t^{(j)} \equiv 0$, this SSFV is the variance of the residual part $\mathcal{Z}_i$. Therefore, when $\beta_t^{(j)} \equiv 0$ for all t, substracting the residual variance $v = \frac{1}{\mathcal{T}}\langle Z, Z\rangle_{\mathcal{T}}$, $V_j - v$ is just the statistical noise, which admits CLT. □

Before stating the first theorem, we define the following quantity:

$$\boldsymbol{\vartheta}_i \triangleq K_n^{-1/2}\left(\vartheta_{i,q+1}, \dots, \vartheta_{i,p}\right)^{\intercal},$$

where

$$\begin{aligned} \vartheta_{i,j} &\triangleq \Delta t_n^{-1}\left(\mathcal{Z}_i^{(0)}\right)^{\intercal}\tilde{\mathcal{X}}_{(j),i}\left(\tilde{\mathcal{X}}_{(j),i}^{\intercal}\tilde{\mathcal{X}}_{(j),i}\right)^{-1}\tilde{\mathcal{X}}_{(j),i}^{\intercal}\mathcal{Z}_i^{(0)} - \langle Z, Z\rangle'_{\tau_{n,i-1}}, \\ \mathcal{Z}_i^{(0)} &\triangleq \left(\sigma^Z_{\tau_{n,i-1}}\Delta W^Z_{t_{n,(i-1)M_n+1}}, \sigma^Z_{\tau_{n,i-1}}\Delta W^Z_{t_{n,(i-1)M_n+2}}, \dots, \sigma^Z_{\tau_{n,i-1}}\Delta W^Z_{t_{n,iM_n}}\right)^{\intercal}. \end{aligned} \tag{3.6}$$

Theorem 1. *Denote* $\mathbf{V} = K_n^{1/2}(V_{q+1} - v, \dots, V_p - v)^{\intercal}$ *and* $v = \frac{1}{\mathcal{T}}\langle Z, Z\rangle_{\mathcal{T}}$. *Define a matrix* $\Sigma_{\mathcal{T}}$, *where the* (j,k)*-th element of* $\Sigma_{\mathcal{T}}$ *can be expressed as:*

$$\Sigma_{\mathcal{T}}^{(j,k)} = \frac{2}{\mathcal{T}}\int_0^{\mathcal{T}}\frac{\left(\omega_t^{(j,k)}\right)^2}{\omega_t^{(j,j)}\omega_t^{(k,k)}}\left(\langle Z, Z\rangle'_t\right)^2 dt \textit{ for } q+1 \le j,k \le p,$$

with $\omega_t^{(j,k)} = c_t^{(j,k)} - \left[c_t^{a,(j)}\right]^\intercal \left(c_t^{a,a}\right)^{-1} c_t^{a,(k)}$, $c_t^{a,(j)} = \left(c_t^{(j,1)}, c_t^{(j,2)}, \dots, c_t^{(j,q)}\right)^\intercal$ *and* $c_t^{a,a} = \left\{c_t^{(r,s)}\right\}_{1\le r,s\le q}$. *Under the null hypothesis* H_0 *in* (3.2) *and assuming that* $q < M_n$, p *can be much larger than* M_n, *and* $(p-q)^2/K_n = o(1), M_n = O(1)$ *holds for all* $p > M_n$. *Suppose that* $\sup_{0\le t\le \mathcal{T}} \|c_t\|_{\max} < \infty$ *and there exists some constant* $C_0 > 0$ *such that* $\left\|c_t^{(j,k)} - c_s^{(j,k)}\right\|_2 \le C_0 |t-s|^{1/2}$ *uniformly for all* $1 \le j,k \le p$ *and* $0 \le s < t \le \mathcal{T}$. *Suppose that* $\Sigma_{\mathcal{T}}$ *is measurable with respect to* $\mathcal{F}_t$ *for all* t, *there exist a constant* $c_\Sigma > 0$ *such that* $\lambda_{\min}(\Sigma_{\mathcal{T}}) > c_\Sigma$. *Under Assumption 1, for arbitrary* $\mathbf{u} = (u_{q+1},\dots,u_p)^\intercal \in \mathbb{R}^{p-q}$ *satisfying* $\|\mathbf{u}\| = 1$, *suppose that for some constant* $\delta > 0$,

$$\left(\mathbf{u}^\intercal \Sigma_{\mathcal{T}} \mathbf{u}\right)^{-\frac{2+\delta}{2}} \sum_{i=1}^{K_n} E\left[\left(\mathbf{u}^\intercal \boldsymbol{\vartheta}_i\right)^{2+\delta} |\mathcal{F}_{\tau_{n,i-1}}\right] = o_p(1), \tag{3.7}$$

then $\left(\mathbf{u}^\intercal \Sigma_{\mathcal{T}} \mathbf{u}\right)^{-1/2} (\mathbf{u}^\intercal \mathbf{V})$ *converges stably to a standard normal distribution as* $n \to \infty$.

Proof. The proof of this theorem is collected in Appendix A of the online supplementary material. □

Note that the stable convergence in law is stronger than the convergence in law in usual sense, see, e.g., Jacod and Protter (2011, pp.47) and Mykland and Zhang (2012, pp. 150) for detailed definition and more discussions. The assumption $(p-q)^2/K_n = o(1)$ means that the dimension of the factor processes should be a smaller order of $n^{1/2}$. When the dimension of X gets larger, there would be a non-negligible bias term in our test statistics due to the difference between the discretized method and the continuous process. How to estimate this bias term or construct another efficient test procedure under ultra-high dimension high frequency factor model deserves some further studies.

To estimate the quantity $\upsilon = \frac{1}{\mathcal{T}} \langle Z, Z\rangle_{\mathcal{T}}$, we propose the following estimator:

$$\hat{\upsilon} = \frac{1}{K_n} \sum_{i=1}^{K_n} \hat{\upsilon}_i \text{ with } \hat{\upsilon}_i = \frac{\mathcal{Y}_i^\intercal \left(\mathbb{I}_{M_n} - \mathcal{H}_{a,i}\right) \mathcal{Y}_i}{\Delta t_n \left(M_n - q\right)}. \tag{3.8}$$

Define the following quantity:

$$\hat{\boldsymbol{\vartheta}}_i \triangleq K_n^{-1/2} \left(\hat{\vartheta}_{i,q+1}, \dots, \hat{\vartheta}_{i,p}\right)^\intercal,$$

where

$$\hat{\vartheta}_{i,j} \triangleq \Delta t_n^{-1} \left(\mathcal{Z}_i^{(0)}\right)^\intercal \left[\tilde{\mathcal{X}}_{(j),i} \left(\tilde{\mathcal{X}}_{(j),i}^\intercal \tilde{\mathcal{X}}_{(j),i}\right)^{-1} \tilde{\mathcal{X}}_{(j),i}^\intercal - \frac{\mathbb{I}_{M_n} - \mathcal{H}_{a,i}}{M_n - q}\right] \mathcal{Z}_i^{(0)},$$

and $\mathcal{Z}_i^{(0)}$ is defined in (3.6).

By replacing the limit υ with $\hat{\upsilon}$ in Theorem 1 , we obtain the following proposition.

Proposition 1. *Denote* $\hat{\mathbf{V}} = K_n^{1/2}(V_{q+1} - \hat{\upsilon}, \dots, V_p - \hat{\upsilon})^\intercal$. *Define a matrix* $\Psi_{\mathcal{T}}$, *where the* (j,k)*-th element of* $\Psi_{\mathcal{T}}$ *can be expressed as* $\Psi_{\mathcal{T}}^{(j,k)} = \Sigma_{\mathcal{T}}^{(j,k)} - \frac{1}{M_n - q} \Sigma_{\mathcal{T}}^{(q+1,q+1)}$ *for* $q+1 \le j,k \le p$. *Under the null hypothesis* H_0 *in* (3.2) *and assuming the same conditions as Theorem 1, also assuming that* $\lambda_{\min}(\Sigma_{\mathcal{T}}) > \frac{1}{M_n - q} \Sigma_{\mathcal{T}}^{(q+1,q+1)}$, *then for arbitrary* $\mathbf{u} = (u_{q+1},\dots,u_p)^\intercal \in \mathbb{R}^{p-q}$ *satisfying* $\|\mathbf{u}\| = 1$, *suppose that for some constant* $\delta > 0$,

$$\left(\mathbf{u}^\intercal \Psi_{\mathcal{T}} \mathbf{u}\right)^{-\frac{2+\delta}{2}} \sum_{i=1}^{K_n} E\left[\left(\mathbf{u}^\intercal \hat{\boldsymbol{\vartheta}}_i\right)^{2+\delta} |\mathcal{F}_{\tau_{n,i-1}}\right] = o_p(1), \tag{3.9}$$

we have: $\left(\mathbf{u}^\intercal \Psi_{\mathcal{T}} \mathbf{u}\right)^{-1/2} \left(\mathbf{u}^\intercal \hat{\mathbf{V}}\right)$ *converges stably to a standard normal distribution as* $n \to \infty$.

Proof. The proof of this proposition is collected in Appendix B of the online supplementary material. □

Remark 3. Note that the Lyapunov-type conditions in (3.7) and (3.9) allow the largest eigenvalues of $\Sigma_{\mathcal{T}}$ and $\Psi_{\mathcal{T}}$, i.e., $\lambda_{\max}\left(\Sigma_{\mathcal{T}}\right)$ and $\lambda_{\max}\left(\Psi_{\mathcal{T}}\right)$, diverge as $p - q$ goes to infinity. The application of our methodologies in finance can be greatly benefited by this theoretical assumption. For example, i*n financial market, the asset prices are usually expressed by a factor model, which allowing for the co-movements among the asset prices. Then, the covariance matrix can have a low-rank plus sparse structure, see, i.e., Aït-Sahalia and Xiu (2017), Fan et al. (2016), Fan and Kim (2018), Kim et al. (2018) and Kong (2018). Therefore, it is necessary to allow* $\lambda_{\max}\left(\Sigma_{\mathcal{T}}\right)$ and $\lambda_{\max}\left(\Psi_{\mathcal{T}}\right)$ diverge.

Remark 4. The main idea to approximate the quantities defined on continuous time is similar to the approach of Riemann sum in approximating some integral. For example, we have divided the sampling period $[0, \mathcal{T}]$ into K_n shrinking time intervals, i.e., $[\tau_{n,i-1}, \tau_{n,i})$ for $i = 1, 2, \dots, K_n$. On each shrinking interval $[\tau_{n,i-1}, \tau_{n,i})$, we can assume the volatility process and co-volatility process to be a constant (e.g., take the value at time $\tau_{n,i-1}$), and thus, we can construct the test statistics locally, e.g., $V_{j,i}$ and $\hat{\upsilon}_i$. To obtain the continuous time version, we can simply aggregate the local test statistics, which is analogous to the construction of Riemann sum, i.e., $V_j = \frac{1}{K_n} \sum_{i=1}^{K_n} V_{j,i}$ and $\hat{\upsilon} = \frac{1}{K_n} \sum_{i=1}^{K_n} \hat{\upsilon}_i$. Since $\mathcal{T}$ is fixed and the block size $\Delta\tau_n = \tau_{n,i} - \tau_{n,i-1}$ is shrinking in asymptotics, we know that the discretization error of the aggregated values, i.e., V_j and $\hat{\upsilon}$, will be negligible as $n \to \infty$.

Based on the idea of observed AVAR proposed in Mykland and Zhang (2017), we provide the nonparametric estimator for the asymptotic covariance matrix $\Psi_{\mathcal{T}}$ as follows:

$$\hat{\Psi}_{\mathcal{T}} = \frac{1}{2K_n} \sum_{i=1}^{K_n-1} \left(D_{i+1} - D_i\right) \left(D_{i+1} - D_i\right)^{\intercal}, \tag{3.10}$$

where $D_i = (V_{q+1,i} - \hat{v}_i, \dots, V_{p,i} - \hat{v}_i)^{\intercal}$.

The properties of $\hat{\Psi}_{\mathcal{T}}$ are stated in the following proposition.

Proposition 2. *Under the same condition of Proposition 1, we have:*

1. *For arbitrary $\mathbf{u} = (u_{q+1}, \dots, u_p)^{\intercal} \in \mathbb{R}^{p-q}$ satisfying $\|\mathbf{u}\| = 1$, we have $\mathbf{u}^{\intercal}\hat{\Psi}_{\mathcal{T}}\mathbf{u} = \mathbf{u}^{\intercal}\Psi_{\mathcal{T}}\mathbf{u} + o_p(1)$ which implies the consistency of $\hat{\Psi}_{\mathcal{T}}$;*
2. *$\hat{\Psi}_{\mathcal{T}}$ is positive definite with probability approaching 1.*

Proof. The proof of this proposition is collected in Appendix C of the online supplementary material. □

3.1. Sum-type test

Based on the results in Proposition 1, we propose the sum-type test statistic for the hypothesis in (3.2) as follows:

$$T_{\mathrm{SUM}} \triangleq \frac{K_n^{1/2}\left(\sum_{j=q+1}^{p} V_j - (p-q)\,\hat{v}\right)}{\sqrt{\mathbf{1}_{(p-q)}^{\intercal}\hat{\Psi}_{\mathcal{T}}\mathbf{1}_{(p-q)}}} \tag{3.11}$$

where $\mathbf{1}_{(p-q)} = (1, 1, \dots, 1)^{\intercal}$ is a $(p-q) \times 1$ column vector with all elements as one.

The asymptotic null distribution of T_{SUM} is obtained in the following proposition.

Proposition 3. *Under the null hypothesis H_0 in (3.2) and assuming the same conditions in Proposition 1, as $n \to \infty$ and $p - q \to \infty$, we have:*

$$T_{SUM} \xrightarrow{\mathcal{L}} N(0,1) \text{ stably in law.}$$

Proof. The proof of this proposition is collected in Appendix D of the online supplementary material. □

Before stating the results about the power of the sum-type test statistic T_{SUM}, we first define several related quantities. Define the correlation matrix $\rho_V = \left\{\rho_V^{(j,k)}\right\}_{q+1\le j,k\le p}$ as

$$\rho_V = \mathrm{Diag}\left(\Psi_{\mathcal{T}}\right)^{-1/2}\Psi_{\mathcal{T}}\mathrm{Diag}\left(\Psi_{\mathcal{T}}\right)^{-1/2}, \tag{3.12}$$

where $\mathrm{Diag}(W)$ denotes the diagonal matrix of W, and $\Psi_{\mathcal{T}}$ is defined in Proposition 1. Define

$$\hat{\psi}^2 = \frac{1}{p-q}\sum_{j=q+1}^{p} \hat{\Psi}_{\mathcal{T}}^{(j,j)}, \tag{3.13}$$

and for $q + 1 \le j, k \le p$, define

$$\hat{\rho}_V^{(j,k)} = \hat{\Psi}_{\mathcal{T}}^{(j,k)}/\hat{\psi}^2. \tag{3.14}$$

Based on the similar arguments in the proof of Proposition 2, we know that $\sup_{q+1\le j\le p}\left|\hat{\psi}^2 - \Psi_{\mathcal{T}}^{(j,j)}\right| = O_p\left(K_n^{-1/2}\right)$, and consequently, for arbitrary $\mathbf{u} = (u_{q+1}, \dots, u_p)^{\intercal} \in \mathbb{R}^{p-q}$ satisfying $\|\mathbf{u}\| = 1$, we have $\mathbf{u}^{\intercal}\hat{\rho}_V\mathbf{u} = \mathbf{u}^{\intercal}\rho_V\mathbf{u} + o_p(1)$.

Define

$$\phi_j \triangleq \frac{K_n^{1/2}\left(V_j - \hat{v}\right)}{\hat{\psi}} \text{ and } \hat{\phi} \triangleq \left(\phi_{q+1}, \phi_{q+2}, \dots, \phi_p\right). \tag{3.15}$$

Then under the null hypothesis H_0 and the same conditions as Proposition 1, we have: for arbitrary $\mathbf{u} = (u_{q+1}, \dots, u_p)^{\intercal} \in \mathbb{R}^{p-q}$ satisfying $\|\mathbf{u}\| = 1$, as $n \to \infty$,

$$\left(\mathbf{u}^{\intercal}\rho_V\mathbf{u}\right)^{-1/2}\left(\mathbf{u}^{\intercal}\hat{\phi}\right) \text{ converges stably to a standard normal distribution.} \tag{3.16}$$

where ρ_V is defined in formula (3.12).

We consider the following alternative hypothesis:

$$H_1 : \beta_t^{(k)} \neq 0, k \in \mathcal{M}, |\mathcal{M}| = m \ge 1, q+1 \le k \le p. \tag{3.17}$$

Next, we will state the performance of ϕ_j under the alternative hypothesis H_1 as stated in (3.17).

Proposition 4. *Under the alternative hypothesis H_1 as stated in (3.17) and assuming the same conditions in Proposition 1, as $n \to \infty$, we have:*

$$\phi_j = K_n^{1/2} \frac{M_n}{\hat{\psi}\mathcal{T}} \int_0^{\mathcal{T}} \boldsymbol{\beta}_{\mathcal{M},t}^{\intercal} \mathcal{P}_t^{\mathcal{M},a,(j),i} \boldsymbol{\beta}_{\mathcal{M},t} dt + O_p(1), \tag{3.18}$$

where $\{j_1, \dots, j_m\} = \mathcal{M}, \boldsymbol{\beta}_{\mathcal{M},t} = \left(\beta_t^{(j_1)}, \dots, \beta_t^{(j_m)}\right)^{\intercal}$,

$$\mathcal{P}_t^{\mathcal{M},a,(j),i} = \left(\omega_t^{(j,j)}\right)^{-1} \zeta_t^{\mathcal{M},a,(j)} \left(\zeta_t^{\mathcal{M},a,(j)}\right)^{\intercal} - \frac{1}{M_n - q} \left[c_t^{\mathcal{M},\mathcal{M}} - \left(c_t^{\mathcal{M},a}\right) \left(c_t^{a,a}\right)^{-1} \left(c_t^{\mathcal{M},a}\right)^{\intercal}\right]$$

with $\omega_t^{(j,j)}, c_t^{a,(j)}$ *and* $c_t^{a,a}$ *being defined in Theorem 1, and*

$$\begin{aligned} c_t^{\mathcal{M},\mathcal{M}} &= \left(c_t^{(j_r,j_s)}\right)_{1 \le r,s \le m}, c_t^{\mathcal{M},(j)} = \left(c_t^{(j,j_1)}, c_t^{(j,j_2)}, \dots, c_t^{(j,j_m)}\right)^{\intercal}, \\ c_t^{\mathcal{M},a} &= \left(c_t^{(j_r,s)}\right)_{1 \le r \le m, 1 \le s \le q}, \zeta_t^{\mathcal{M},a,(j)} = c_t^{\mathcal{M},(j)} - c_t^{\mathcal{M},a} \left(c_t^{a,a}\right)^{-1} c_t^{a,(j)}. \end{aligned}$$

Proof. The proof of this proposition is collected in Appendix E of the online supplementary material. □

Remark 5 (*Discussion on Signal Strength*)**.** Before stating the results about signal strength, we first define some useful quantities. For the group of variables with non-zero beta, we define $X_{\mathcal{M},t} = \left(X_t^{(j_1)}, \dots, X_t^{(j_m)}\right)^{\intercal}$, and $\mathcal{X}_{\mathcal{M},i} = \left(\Delta X_{\mathcal{M},t_{n,(i-1)M_n+1}}, \Delta X_{\mathcal{M},t_{n,(i-1)M_n+2}}, \dots, \Delta X_{\mathcal{M},t_{n,iM_n}}\right)^{\intercal}$. We define the set of variables with non-zero regression coefficients as *group $\mathcal{M}$ variables*. Then based on the proof of Proposition 4, we know that under H_1, the dominating term in $V_{j,i} - \hat{v}_i$ will be:

$$\underbrace{\boldsymbol{\beta}_{\mathcal{M},\tau_{n,i-1}}^{\intercal} \mathcal{X}_{\mathcal{M},i}^{\intercal} \tilde{\mathcal{X}}_{(j),i} \left(\tilde{\mathcal{X}}_{(j),i}^{\intercal} \tilde{\mathcal{X}}_{(j),i}\right)^{-1} \tilde{\mathcal{X}}_{(j),i}^{\intercal} \mathcal{X}_{\mathcal{M},i} \boldsymbol{\beta}_{\mathcal{M},\tau_{n,i-1}}}_{Q_{(j),i}} - \frac{1}{M_n - q} \underbrace{\boldsymbol{\beta}_{\mathcal{M},\tau_{n,i-1}}^{\intercal} \mathcal{X}_{\mathcal{M},i}^{\intercal} \left(\mathbb{I}_{M_n} - \mathcal{H}_{a,i}\right) \mathcal{X}_{\mathcal{M},i} \boldsymbol{\beta}_{\mathcal{M},\tau_{n,i-1}}}_{Q_{\mathcal{M},i}},$$

where by further defining the component of group $\mathcal{M}$ variables which is orthogonal to the group a variables:

$$\tilde{\mathcal{X}}_{\mathcal{M},i} = \left(\mathbb{I}_{M_n} - \mathcal{H}_{a,i}\right) \mathcal{X}_{\mathcal{M},i},$$

we have:

$$\begin{aligned} Q_{(j),i} &= \left\|\tilde{\mathcal{X}}_{(j),i}\right\|^{-2} \left\|\tilde{\mathcal{X}}_{(j),i}^{\intercal} \tilde{\mathcal{X}}_{\mathcal{M},i} \boldsymbol{\beta}_{\mathcal{M},\tau_{n,i-1}}\right\|^2 \le \left\|\tilde{\mathcal{X}}_{\mathcal{M},i}\right\|^2 \left\|\boldsymbol{\beta}_{\mathcal{M},\tau_{n,i-1}}\right\|^2, \\ Q_{\mathcal{M},i} &= \left\|\tilde{\mathcal{X}}_{\mathcal{M},i} \boldsymbol{\beta}_{\mathcal{M},\tau_{n,i-1}}\right\|^2 \le \left\|\tilde{\mathcal{X}}_{\mathcal{M},i}\right\|^2 \left\|\boldsymbol{\beta}_{\mathcal{M},\tau_{n,i-1}}\right\|^2, \end{aligned}$$

where $\|\cdot\|$ denotes the Euclidean norm (resp. spectral norm) for vector (resp. matrix).

Therefore, based on the two inequalities above, it is easy to see that the two main sources which affecting the signal strength of $V_{j,i} - \hat{v}_i$ are $\left\|\boldsymbol{\beta}_{\mathcal{M},\tau_{n,i-1}}\right\|$ and $\|\tilde{\mathcal{X}}_{\mathcal{M},i}\|$. The detailed discussion about their impact is as follows. First, higher $\left\|\boldsymbol{\beta}_{\mathcal{M},\tau_{n,i-1}}\right\|$ yields stronger signal in $V_{j,i} - \hat{v}_i$; lower $\left\|\boldsymbol{\beta}_{\mathcal{M},\tau_{n,i-1}}\right\|$ yields weaker signal in $V_{j,i} - \hat{v}_i$. Consequently, we know that $\mathcal{T}^{-1} \int_0^{\mathcal{T}} \|\boldsymbol{\beta}_{\mathcal{M},t}\| dt$ has positive relationship with the signal strength in $V_j - \hat{v}$. Second, higher $\|\tilde{\mathcal{X}}_{\mathcal{M},i}\|$ yields stronger signal in $V_{j,i} - \hat{v}_i$; lower $\|\tilde{\mathcal{X}}_{\mathcal{M},i}\|$ yields weaker signal in $V_{j,i} - \hat{v}_i$.

We could further explain the influencer of $\|\tilde{\mathcal{X}}_{\mathcal{M},i}\|$ from a geometric perspective. Note that $\tilde{\mathcal{X}}_{\mathcal{M},i}$ is the components of the group $\mathcal{M}$ variables which is orthogonal to the group a variables. Then it is natural that if the original directions of the group $\mathcal{M}$ variables are closer to the directions of group a variables, the magnitude of the orthogonal components $\tilde{\mathcal{X}}_{\mathcal{M},i}$ will be smaller, which leads to smaller $\|\tilde{\mathcal{X}}_{\mathcal{M},i}\|$. From a statistical perspective, it means that if the group a variables have higher correlation with the group $\mathcal{M}$ variables, $\|\tilde{\mathcal{X}}_{\mathcal{M},i}\|$ becomes smaller. In the extreme case, if the group $\mathcal{M}$ variables is a linear combination of the group a variables, then $\|\tilde{\mathcal{X}}_{\mathcal{M},i}\| = 0$, and in this case, the power of the proposed tests will be low. Therefore, throughout this paper, when we are discussing the theoretical results about the behavior of our tests under the alternative hypothesis, we assume that the absolute value of the correlation between group $\mathcal{M}$ variables and group a variables are bounded away from 1 in asymptotics. □

Remark 6. When $|\mathcal{M}| = 1$, and $\beta_t^{(j)} \neq 0$, then we have:

$$\begin{aligned} \mathcal{P}_t^{\mathcal{M},a,(j),i} &= \left(1 - \frac{1}{M_n - q}\right) \omega_t^{(j,j)}, \\ \mathcal{P}_t^{\mathcal{M},a,(k),i} &= \left(\omega_t^{(k,k)}\right)^{-1} \left[\omega_t^{(j,k)}\right]^2 - \frac{1}{M_n - q} \omega_t^{(j,j)}, \text{ for } k \neq j \end{aligned}$$

and therefore, we have:

$$\int_0^{\mathcal{T}} \boldsymbol{\beta}_{\mathcal{M},t}^{\intercal} \mathcal{P}_t^{\mathcal{M},a,(j),i} \boldsymbol{\beta}_{\mathcal{M},t} dt = \left(1 - \frac{1}{M_n - q}\right) \int_0^{\mathcal{T}} \omega_t^{(j,j)} \left[\beta_t^{(j)}\right]^2 dt, \tag{3.19}$$

and for $k \neq j$,

$$\int_0^{\mathcal{T}} \boldsymbol{\beta}_{\mathcal{M},t}^{\intercal} \mathcal{P}_t^{\mathcal{M},a,(k),i} \boldsymbol{\beta}_{\mathcal{M},t} dt = \int_0^{\mathcal{T}} \left(\omega_t^{(k,k)}\right)^{-1} \left[\omega_t^{(j,k)}\right]^2 \left[\beta_t^{(j)}\right]^2 dt - \frac{1}{M_n - q} \int_0^{\mathcal{T}} \omega_t^{(j,j)} \left[\beta_t^{(j)}\right]^2 dt.$$

where $\omega_t^{(j,k)}$ is defined in Theorem 1. For $|\mathcal{M}| > 1$, by assuming that

$$\sup_{0\le t\le \mathcal{T}} \left\| \mathcal{P}_t^{\mathcal{M},a,(j),i} \right\|_{\max} = O_p(1), \tag{3.20}$$

then we have:

$$\int_0^{\mathcal{T}} \beta_{\mathcal{M},t}^{\intercal} \mathcal{P}_t^{\mathcal{M},a,(j),i} \beta_{\mathcal{M},t} dt = O_p\left(|\mathcal{M}|^2\right). \tag{3.21}$$

On the other hand, by detailed but tedious derivation, we have:

$$\hat{\psi}^2 \asymp |\mathcal{M}|^4 \ \text{ for } q+1 \le j \le p \text{ and } |\mathcal{M}| \ge 1. \tag{3.22}$$

Finally, based on (3.18), (3.21), and (3.22) , suppose (i) the absolute value of the correlation between group $\mathcal{M}$ variables and group a variables are bounded away from 1 in asymptotics; (ii) there exists some constant $c_{\mathcal{M}} > 0$, such that $\mathcal{T}^{-1}\int_0^{\mathcal{T}} \|\beta_{\mathcal{M},t}\|\, dt > c_{\mathcal{M}}$, then we know that

$$\phi_j \asymp K_n^{1/2} \text{ for } q+1 \le j \le p \text{ and } |\mathcal{M}| \ge 1. \quad \square$$

Proposition 5 (*Power of Sum-Type Test*). *Suppose (i) the absolute value of the correlation between group $\mathcal{M}$ variables and group a variables are bounded away from 1 in asymptotics; (ii) there exists some constant $c_{\mathcal{M}} > 0$, such that $\mathcal{T}^{-1}\int_0^{\mathcal{T}} \|\beta_{\mathcal{M},t}\|\, dt > c_{\mathcal{M}}$, then we have $T_{SUM} \asymp K_n^{1/2}$ which means that T_{SUM} will explode under H_1. Thus, the type-II error of the sum-type test is asymptotically negligible.*

Proof. The proof of this proposition is based on the discussions in Remarks 5 and 6. □

3.2. Max-type test

We propose the max-type test statistic as follows:

$$T_{\mathrm{MAX}} \triangleq K_n \max_{q+1\le j\le p} \frac{\left(V_j - \hat{v}\right)^2}{\hat{\psi}^2}, \tag{3.23}$$

where $\hat{\psi}^2$ is defined in (3.13).

Before stating the limiting null distribution of the max-type test statistic, we introduce a necessary assumption for the correlation matrix ρ_V as follows.

Assumption 2. For some $\varrho \in (0,1)$, assume $\left|\rho_V^{(j,k)}\right| \le \varrho$ for all $q+1 \le j < k \le p$. Suppose $\left\{\delta_{p-q} : p-q \ge 1\right\}$ and $\left\{\varsigma_{p-q} : p-q \ge 1\right\}$ are positive constants with $\delta_{p-q} = o(1/\log(p-q))$ and $\varsigma = \varsigma_{p-q} \to 0$ as $p-q \to \infty$. For $q+1 \le j \le p$, define $B_{p,q,j} = \left\{q+1 \le k \le p : \left|\rho_V^{(j,k)}\right| > \delta_{p-q}\right\}$ and $C_{p,q} = \left\{q+1 \le j \le p : \left|B_{p,q,j}\right| > (p-q)^{\varsigma}\right\}$. We assume that $\left|C_{p,q}\right|/(p-q) \to 0$ as $p-q \to \infty$.

Assumption 2 means that, for each variable, there should not be too many variables which are strongly correlated with it. For example, if ρ_V is a banded correlation matrix, i.e. $\rho_V^{(j,k)} = 0$ if $|j-k| > L$ with fixed $L > 0$, then $|C_{p-q}| = 0$ and Assumption 2 holds. The asymptotic null distribution of the max-type test statistic can be expressed as follows.

Proposition 6. *Suppose Assumption 2. Assume the same conditions in Proposition 1. Then, under the null hypothesis H_0 in (3.2), for any $y \in \mathbb{R}$, as $n \to \infty$ and $p-q \to \infty$, we have:*

$$\left|P(T_{MAX} - 2\log(p-q) + \log\log(p-q) \le y) - \exp\left\{-\pi^{-1/2}\exp(-y/2)\right\}\right| = o(1).$$

Proof. The proof of this proposition is collected in Appendix F of the online supplementary material. □

Proposition 7 (*Power of MAX-Type Test*). *Suppose the same conditions in Proposition 5, under the alternative hypothesis H_1 as stated in (3.17), we have $T_{MAX} \asymp K_n$. Therefore, because $\log(p-q)/K_n = o\left((p-q)^2/K_n\right)$, under the assumption $(p-q)^2/K_n = o(1)$, we know that T_{MAX} will explode and in this case the type-II error of the max-type test is asymptotically negligible.*

3.3. Asymptotic independence and Fisher's combination test

Intuitively speaking, the sum-type test procedure sums the signals of each variable and then performs better under the dense alternatives, i.e., there are many small non-zero signals. If there are only a few large signals (sparse alternative), the signals would be dominated by the variance of sum-type test statistics. Therefore, the sum-type test procedure does not perform well under sparse alternative scenarios. In contrast, the max-type test procedure only considers the maximum of the signals. Therefore, if there are very few large signals, the max-type test procedure would catch the largest signal and perform better. However, for the dense alternative scenario, the maximum of many small signals would also be very small. Therefore, the max-type test procedure cannot have any power under this case.

In the existing statistical literature on high dimensional mean vector test with low frequency data, the sum-type test can only work well with the dense alternative while the max-type test can only work well with the sparse alternative. To achieve good performance under both sparse and dense alternatives, statisticians usually combine the sum-type test and max-type test based on their asymptotic independence. In this section, we first show the asymptotic independence between our sum-type test and max-type test and then derive the Fisher's combination test.

Before stating the asymptotic independence between sum-type test and max-type test, we first impose several assumptions for the correlation matrix ρ_V.

Assumption 3. Suppose $\lambda_{\min}\left(\rho_V\right) > c_\rho$ for some constant $c_\rho > 0$. Suppose that there exists $C_R > 0$, so that $\max_{q+1\le j\le p}\sum_{k=q+1}^{p}\left|\rho_V^{(j,k)}\right| \le C_R$ as $p-q\to\infty$.

Remark 7. Note that Assumption 3 is stronger than Assumption 2, which can be shown as follows. Take $\delta_{p-q} = 1/(\log(p-q))^2$. Recall $B_{p,q,j} = \left\{q+1\le k\le p : \left|\rho_V^{(j,k)}\right| > \delta_{p-q}\right\}$ defined in Assumption 2. By Assumption 3, we know that

$$\left|B_{p,q,j}\right|\cdot\frac{1}{(\log(p-q))^2} \le \sum_{k=q+1}^{p}\left|\rho_V^{(j,k)}\right| \le C_R$$

for each $q+1 \le j \le p$. This shows that $\max_{q+1\le j\le p}\left|B_{p,q,j}\right| \le C_R(\log(p-q))^2$. Take $\varsigma = \varsigma_{p-q} = \left(\log_{\log(p-q)} C_R + 3\right)\log\log(p-q)/\log(p-q)$ for $p-q \ge e^e$. Then $\varsigma_{p-q} \to 0$ and $C_R(\log(p-q))^2 < (p-q)^\varsigma$. As a result, $C_{p,q} = \left\{q+1\le j\le p : \left|B_{p,q,j}\right| > (p-q)^\varsigma\right\} = \varnothing$. So the sparsity assumption of the correlation matrix ρ_V in Assumption 3 is more restrictive than Assumption 2, which requires the sum of the absolute values of correlation between each variable with the other variables are smaller than a positive constant. Specially, we can also show that the banded correlation matrix with a fixed band length L also satisfies Assumption 3.

The asymptotic independence between sum-type test and max-type test can be stated as follows.

Theorem 2. *Suppose Assumption 3. Assume the same conditions in Proposition 1. Then, under the null hypothesis H_0 in (3.2), for any $x, y \in \mathbb{R}$, we have:*

$$P\left(T_{SUM} \le x, T_{MAX} - 2\log(p-q) + \log\log(p-q) \le y\right) \to \Phi(x)\,F(y),$$

where $\Phi(x)$ denotes the cumulative distribution function of standard normal distribution and $F(y) = \exp\left\{-\pi^{-1/2}\exp(-y/2)\right\}$.

Proof. The proof of this theorem is collected in Appendix G of the online supplementary material. □

To combine the proposed sum-type and max-type tests, we propose the Fisher's combination test, based on the asymptotic independence between T_{SUM} and T_{MAX}. Specifically, let

$$p_{\mathrm{SUM}} \triangleq 1 - \Phi\left(T_{\mathrm{SUM}}\right)$$

and

$$p_{\mathrm{MAX}} \triangleq 1 - F\left(T_{\mathrm{MAX}} - 2\log(p-q) + \log\log(p-q)\right)$$

denote the p-values with respect to the test statistics T_{SUM} and T_{MAX}, respectively. Based on p_{SUM} and $p_{\ \mathrm{MAX}}$, the proposed Fisher's combination test rejects H_0 at the significance level α, if

$$T_{\mathrm{FC}} \triangleq -2\log p_{\mathrm{SUM}} - 2\log p_{\mathrm{MAX}} \tag{3.24}$$

is larger than the $1-\alpha$ quantile of the chi-squared distribution with 4 degrees of freedom.

Based on Theorem 2, we have the following result for T_{FC}.

Proposition 8. *Assume the same conditions as in Theorem 2, then we have $T_{FC} \xrightarrow{\mathcal{L}} \chi_4^2$ as $n\to\infty$ and $p-q\to\infty$.*

Proof. The validity of this proposition can be easily verified based on the results in Littell and Folks (1971, 1973). □

Proposition 9 (*Power of Fisher's Combination Test*)**.** *Under the assumption $(p-q)^2/K_n = o(1)$ and the same conditions in Propositions 5 and 7, we know that both T_{SUM} and $T_{\ MAX}$ will explode as $n\to\infty$ and $p-q\to\infty$ and consequently, we have $p_{SUM}\to 0$ and $p_{MAX}\to 0$ in asymptotics. In this case, it is easy to see that T_{FC} explodes and thus the type-II error of the Fisher's combination test is asymptotically negligible.*

3.4. Discussion about jump, market microstructure noise and irregular/asynchronous observation time

In practice, high frequency financial data are often contaminated by the jumps, market microstructure noise and irregular/asynchronous observation time, which may introduce additional problems in the statistical inference. Therefore, when applying the proposed tests to the real-world high frequency data, the testing procedure should be carefully designed to obtain the correct result. In this subsection, we mainly discuss the techniques that can be applied to eliminate the harmful impact of jumps, microstructure noise and irregular/asynchronous observation time.

3.4.1. Jump

Jumps are widely observed in the intraday asset returns. The independent variable process $X_t = \left(X_t^{(1)}, X_t^{(2)}, \ldots, X_t^{(p)}\right)^\top$ and the residual process Z_t can also incorporate jump processes by following the Assumptions 1, 2, and 3 in Aït-Sahalia et al. (2020), i.e.,

$$\begin{aligned} dX_t &= \mu_t^X dt + \sigma_t^X dW_t + dJ_t^X, \\ dZ_t &= \mu_t^Z dt + \sigma_t^Z dW_t^Z + dJ_t^Z, \end{aligned}$$

and the triplet $\left(Y_t, X_t, Z_t\right)$ has the relationship as described in (3.1).

By applying the truncation technique to the observed log returns, our theories that were established for the continuous path still work. More specifically, by replacing $\mathcal{Y}_i$, $\mathcal{X}_{a,i}$, and $\mathcal{X}_{(j),i}$ (as defined in (3.3), (3.4) and (3.5)) with

$$\mathcal{Y}_i = \begin{pmatrix} \Delta Y_{t_{n,(i-1)M_n+1}} \mathbf{1}_{\left\{\left|\Delta Y_{t_{n,(i-1)M_n+1}}\right| \le u_n\right\}} \\ \Delta Y_{t_{n,(i-1)M_n+2}} \mathbf{1}_{\left\{\left|\Delta Y_{t_{n,(i-1)M_n+2}}\right| \le u_n\right\}} \\ \vdots \\ \Delta Y_{t_{n,iM_n}} \mathbf{1}_{\left\{\left|\Delta Y_{t_{n,iM_n}}\right| \le u_n\right\}} \end{pmatrix}_{M_n \times 1}, \mathcal{X}_{a,i} = \begin{pmatrix} \Delta X^\top_{a,t_{n,(i-1)M_n+1}} 1_{\left\{\left\|\Delta X_{a,t_{n,(i-1)M_n+1}}\right\| \le u_n\right\}} \\ \Delta X^\top_{a,t_{n,(i-1)M_n+2}} 1_{\left\{\left\|\Delta X_{a,t_{n,(i-1)M_n+2}}\right\| \le u_n\right\}} \\ \vdots \\ \Delta X^\top_{a,t_{n,iM_n}} 1_{\left\{\left\|\Delta X_{a,t_{n,iM_n}}\right\| \le u_n\right\}} \end{pmatrix}_{M_n \times q},$$

and

$$\mathcal{X}_{(j),i} = \begin{pmatrix} \Delta X^{(j)}_{t_{n,(i-1)M_n+1}} \mathbf{1}_{\left\{\left|\Delta X^{(j)}_{t_{n,(i-1)M_n+1}}\right| \le u_n\right\}} \\ \Delta X^{(j)}_{t_{n,(i-1)M_n+2}} \mathbf{1}_{\left\{\left|\Delta X^{(j)}_{t_{n,(i-1)M_n+2}}\right| \le u_n\right\}} \\ \vdots \\ \Delta X^{(j)}_{t_{n,iM_n}} \mathbf{1}_{\left\{\left|\Delta X^{(j)}_{t_{n,iM_n}}\right| \le u_n\right\}} \end{pmatrix}_{M_n \times 1},$$

respectively, where $u_n \asymp \Delta t_n^{\varpi}$ with $\varpi \in (0, 1/2)$, the main results in Theorems 1–2 and Propositions 1–8 still hold.

3.4.2. Market microstructure noise

Market microstructure noise is widespread in the high frequency financial data. As the sampling frequency increases, the harmful impact of market microstructure noise becomes more devastating, see, e.g., Zhang et al. (2005) and Chen et al. (2020, 2024). Existing techniques that are designed to eliminate the impact of noise include: (i) sparse sampling scheme, see, e.g., Zhang et al. (2005, Section 2.3), Aït-Sahalia and Xiu (2017, 2019) and Aït-Sahalia et al. (2020); (ii) pre-averaging approach as proposed in Jacod et al. (2009); (iii) two scales construction or multiple scales construction, see, e.g., Zhang et al. (2005), Zhang (2006), and Mykland et al. (2019). In practice, we recommend that the readers use the sparse sampling scheme when implementing the proposed tests. For example, when using the sampling interval Δt_n no less than 10 min, the microstructure effects become negligible in the analysis.

3.4.3. Irregular/asynchronous observation time

The observation times in high frequency financial data are spaced irregularly and asynchronously, which can introduce additional challenges in the multivariate data analysis, see, e.g., Zhang (2011) and Mykland et al. (2019). Existing techniques to solve this problem include: (i) previous tick technique proposed by Zhang (2011) ; (ii) refresh times proposed by Barndorff-Nielsen et al. (2011); (iii) generalized synchronization method proposed by Aït-Sahalia et al. (2010); (iv) pre-averaging technique, see, e.g., Mykland et al. (2019). When implementing the proposed tests in practice, readers can employ the sparse sampling scheme plus one of the first three techniques mentioned above (i.e., (i)-(iii)), so that the impact of irregular and asynchronous observation times is mitigated.

4. Monte Carlo evidence

Following the multiple regression model defined in (3.1), we further define:

$$dX_t^{(k)} = \mu_k dt + \sigma_t^{(k)} d\mathcal{W}_t^{(k)} \text{ and } dZ_t = \nu_t d\mathcal{B}_t,$$

where $k = 1, 2, \ldots, p$. Throughout this section, we set $q = 3$.

In this simulations, the first factor of $X_t^{(1)}$ is set as the market factor. Thus, its factor loading $\beta_t^{(1)}$ is positive. Therefore, we simulate the factor loading in the following scheme:

$$d\beta_t^{(k)} = \begin{cases} \tilde{\kappa}_1\left(\tilde{\theta}_1 - \beta_t^{(k)}\right)dt + \tilde{\xi}_1\sqrt{\beta_t^{(k)}}d\tilde{\mathcal{B}}_t^{(k)} & \text{if } k = 1, \\ \tilde{\kappa}_k\left(\tilde{\theta}_k - \beta_t^{(k)}\right)dt + \tilde{\xi}_k d\tilde{\mathcal{B}}_t^{(k)} & \text{if } k \geq 2. \end{cases} \tag{4.1}$$

Let us introduce a p-dimensional standard Brownian motion $\tilde{\mathcal{W}}_t$, where the correlation between $d\mathcal{W}_t^{(k)}$ and $d\tilde{\mathcal{W}}_t^{(k)}$ is ρ_k. The correlation matrix of $d\tilde{\mathcal{W}}_t$ is defined as ρ^{σ}. The volatility processes of $\mathbf{X}$ and Z are simulated as follows:

$$d\left(\sigma_t^{(k)}\right)^2 = \kappa_k\left(\theta_k - \left(\sigma_t^{(k)}\right)^2\right)dt + \eta_k\sigma_t^{(k)}d\tilde{\mathcal{W}}_t^{(k)} \text{ and } dv_t^2 = \kappa^{v}\left(\theta^{v} - v_t^2\right)dt + \eta^{v}v_t d\bar{\mathcal{B}}_t.$$

In the simulation, the setting of all parameters are as follows: $\mu = \left(\mu_1, \ldots, \mu_p\right)$ with $\mu_i \sim U(0.01, 0.08)$, $\tilde{\kappa} = \left(1, \tilde{\kappa}_2, \ldots, \tilde{\kappa}_p\right)$ with $\tilde{\kappa}_i \sim U(2, 8)$ for $2 \leq i \leq p$, $\tilde{\xi} = \left(0.5, \tilde{\xi}_2, \ldots, \tilde{\xi}_p\right)$ with $\tilde{\xi}_i \sim U(0.2, 0.9)$ for $2 \leq i \leq p$, $\tilde{\theta}_1 \sim U[0.25, 1.75]$, $\tilde{\theta}_i \sim U(-0.1, 0.1)$ for $2 \leq i \leq p$, $\kappa = \left(\kappa_1, \ldots, \kappa_p\right)$ with $\kappa_i \sim U(2, 8)$, $\theta = \left(\theta_1, \ldots, \theta_p\right)$ with $\theta_i \sim U(0.01, 0.09)$, $\eta = \left(\eta_1, \ldots, \eta_p\right)$ with $\eta_i = \sqrt{\kappa_i\theta_i u_i}$ and $u_i \sim U(0.1, 0.9)$, $\rho = \left(\rho_1, \ldots, \rho_p\right)$ with $\rho_i \sim U(-0.8, -0.1)$, $\rho^{\sigma} = \operatorname{Diag}(\Pi)^{-1/2}\,\Pi\operatorname{Diag}(\Pi)^{-1/2}$ with $\Pi = \xi^{\intercal}\boldsymbol{\xi}, \xi = \left\{\xi_{i,j}\right\}_{1\leq i,j\leq p}$ where $\xi_{i,j} = \xi_{j,i} \sim U(0.01, 0.79)$ for $1 \leq i < j \leq p$, and $\xi_{i,i} = 1$ for $1 \leq i \leq p$. and $\kappa^{v} = 4$, $\theta^{v} = 0.06$ and $\eta^{v} = 0.3$. The parameter setting in the simulation exercise follows the similar setting as the simulation sections of Fan et al. (2016), Aït-Sahalia and Xiu (2019) and Aït-Sahalia et al. (2020). More specifically, the range of the corresponding parameter is similar to that of the previous literature. Moreover, the selected ranges for the parameters in this simulation exercise are close to the real data characteristics in the financial market, see, e.g., Aït-Sahalia and Kimmel (2007).

The time horizon in the simulation experiment is set as: $\mathcal{T} = 1$ month which consists 21 trading days. We assume that a trading day consists of 6.5 h for open trading. The empirical sizes and powers presented in the following subsections are calculated based on 1000 sample paths.

4.1. Empirical size

To show the empirical sizes of the three proposed tests, we set $\beta_t^{(k)} \equiv 0$ for $4 \leq k \leq p$.

Table 4.1 shows the empirical sizes of the three proposed tests under the settings of different Δt_n and p. It is easy to see that all of the three test statistics can control the sizes very well with different Δt_n and p. This is true even if the number of regressors p is very large, e.g., comparable to $K_n^{1/2}$.

Fig. 4.1 shows the histograms of the three proposed tests with $\Delta t_n = 5$ sec and 2.5 min, $p = 60$ and 100, where the red solid lines denote the benchmarks of the limiting null distributions and the blue dashed lines denote the critical values of the related test statistics. All of the histograms are very close to the benchmarks, demonstrating the validity of our theoretical results under the null hypothesis.

4.2. Power

We show the power of the three proposed tests. We first introduce a positive integer r such that $r > q$ and $\beta_t^{(k)} \equiv 0$ for $r+1 \leq k \leq p$. For non-zeros betas, i.e., $\beta_t^{(k)}$ for $k \leq r$, we generate the beta processes based on the equations in formula (4.1), where the market beta process $\beta_t^{(1)}$ is set to be positive.

Fig. 4.2 shows the power curves of the three proposed tests under different settings of Δt_n, r and p.

From Fig. 4.2-(a) to -(d), it is easy to see that when r is low (i.e., $r \leq 150$), the max-type test has higher power than the sum-type test; when r is high (i.e., $250 \leq r \leq 400$), the sum-type test has higher power. Overall, the power of the Fisher's combination test is always higher than that of the sum-type test and max-type test. This is not surprising because the max-type test can work well under the sparse alternative while the sum-type test can work well under the dense alternative, see, e.g., Cai et al. (2014), Bai and Saranadasa (1996), Srivastava and Du (2008), Srivastava (2009), Srivastava et al. (2013), and Chen and Qin (2010). Moreover, as the combination of the sum-type test and max-type test, the Fisher's combination test is robust to both sparse alternative and dense alternative. Fig. 4.2 -(e) and (f) mimick a large model, where the number of none-zero $\beta^{(k)}$ is much smaller than p, i.e. $r \ll p$. We can see that the Fisher's combination test and max-type test outperform the sum-type test. This superior performances become more pronounced with the sparsity (as p increases in (e)). Overall, all of these six plots show the robustness of the Fisher's combination test, which is consistent with the existing literature, see, e.g., Fan et al. (2015), Xu et al. (2016), and He et al. (2021).

5. Empirical study

In this section, we use our proposed tests to analyze real data in the financial market. We present two examples of real data analysis. In the first example, we further examine the validity of our proposed tests by constructing three high frequency regressions, which corresponding to the dense alternative, sparse alternative and null hypothesis, respectively. In the second example, we demonstrate a possible application of our proposed tests through high frequency factor models.

Table 4.1
Empirical sizes of the three proposed tests.

	T_{SUM}			
	$\Delta t_n = 2.5$ min	$\Delta t_n = 5$ min	$\Delta t_n = 15$ min	$\Delta t_n = 30$ min
$p = 20$	0.046	0.050	0.046	0.045
$p = 40$	0.049	0.052	0.056	0.049
$p = 60$	0.046	0.045	0.058	0.058
$p = 80$	0.036	0.040	0.056	0.049
$p = 100$	0.043	0.053	0.049	0.056
$p = 200$	0.041	0.055	0.057	0.053
$p = 300$	0.042	0.046	0.050	0.045
$p = 400$	0.044	0.043	0.060	0.048
$p = 500$	0.051	0.047	0.050	0.046
	T_{MAX}			
	$\Delta t_n = 2.5$ min	$\Delta t_n = 5$ min	$\Delta t_n = 15$ min	$\Delta t_n = 30$ min
$p = 20$	0.044	0.040	0.052	0.047
$p = 40$	0.054	0.039	0.043	0.050
$p = 60$	0.059	0.056	0.049	0.048
$p = 80$	0.059	0.049	0.045	0.042
$p = 100$	0.067	0.062	0.047	0.053
$p = 200$	0.060	0.062	0.053	0.046
$p = 300$	0.055	0.059	0.044	0.049
$p = 400$	0.049	0.057	0.053	0.054
$p = 500$	0.060	0.056	0.044	0.059
	T_{FC}			
	$\Delta t_n = 2.5$ min	$\Delta t_n = 5$ min	$\Delta t_n = 15$ min	$\Delta t_n = 30$ min
$p = 20$	0.055	0.051	0.042	0.052
$p = 40$	0.058	0.042	0.043	0.044
$p = 60$	0.064	0.070	0.048	0.050
$p = 80$	0.059	0.052	0.045	0.053
$p = 100$	0.068	0.064	0.045	0.047
$p = 200$	0.064	0.054	0.051	0.048
$p = 300$	0.058	0.052	0.041	0.046
$p = 400$	0.051	0.053	0.053	0.049
$p = 500$	0.055	0.061	0.041	0.056

Notes. This table reports the empirical sizes of the three proposed tests, including T_{SUM}, T_{MAX} and T_{FC} which are defined in (3.11), (3.23) and (3.24), respectively.

5.1. *Example of high frequency regression*

To verify the performance of our proposed tests, we conduct the empirical study with the intraday stock prices over the period between Jan 2007 and Dec 2017 (in total 2769 trading days).

For the independent variable processes, we mainly employ the dataset of the 80 most actively traded stocks among the components of S&P 100 Index ($X_t^{\text{SP100-1}}, X_t^{\text{SP100-2}}, \dots, X_t^{\text{SP100-80}}$), of which the intraday stock prices are downloaded from the Trade and Quote (TAQ) database of the New York Stock Exchange (NYSE). We have classified these 80 constituents of S&P 100 Index into different sectors based on the Global Industrial Classification Standard (GICS). There are only 3 stocks classified into the Sector of Utilities and there are 5 stocks which do not belong to any sectors. Without loss of generality, we set $X_t^{\text{SP100-1}}, X_t^{\text{SP100-2}}, X_t^{\text{SP100-3}}, X_t^{\text{SP100-4}}$ and $X_t^{\text{SP100-5}}$ as the 5 stocks which do not belong to any sectors and set $X_t^{\text{SP100-6}}, X_t^{\text{SP100-7}}$ and $X_t^{\text{SP100-8}}$ as the 3 stocks belonging to the Sector of Utilities.

The candidates of the dependent variable process Y_t include the iShares S&P 100 ETF (OEF) and the Sector SDPR ETF on Utilities (XLU), of which the intraday stock prices are downloaded from the TAQ database of NYSE.

In this empirical study, we employ the 15-min subsamples for both X_t and Y_t, which implies that $\Delta t_n = 15$ min. The test window is on a monthly basis, i.e., $\mathcal{T} = 1$ month, which consists 21 trading days. For each trading day, the intraday stock prices are collected between 9:35 a.m. EST and 3:55 p.m. EST.

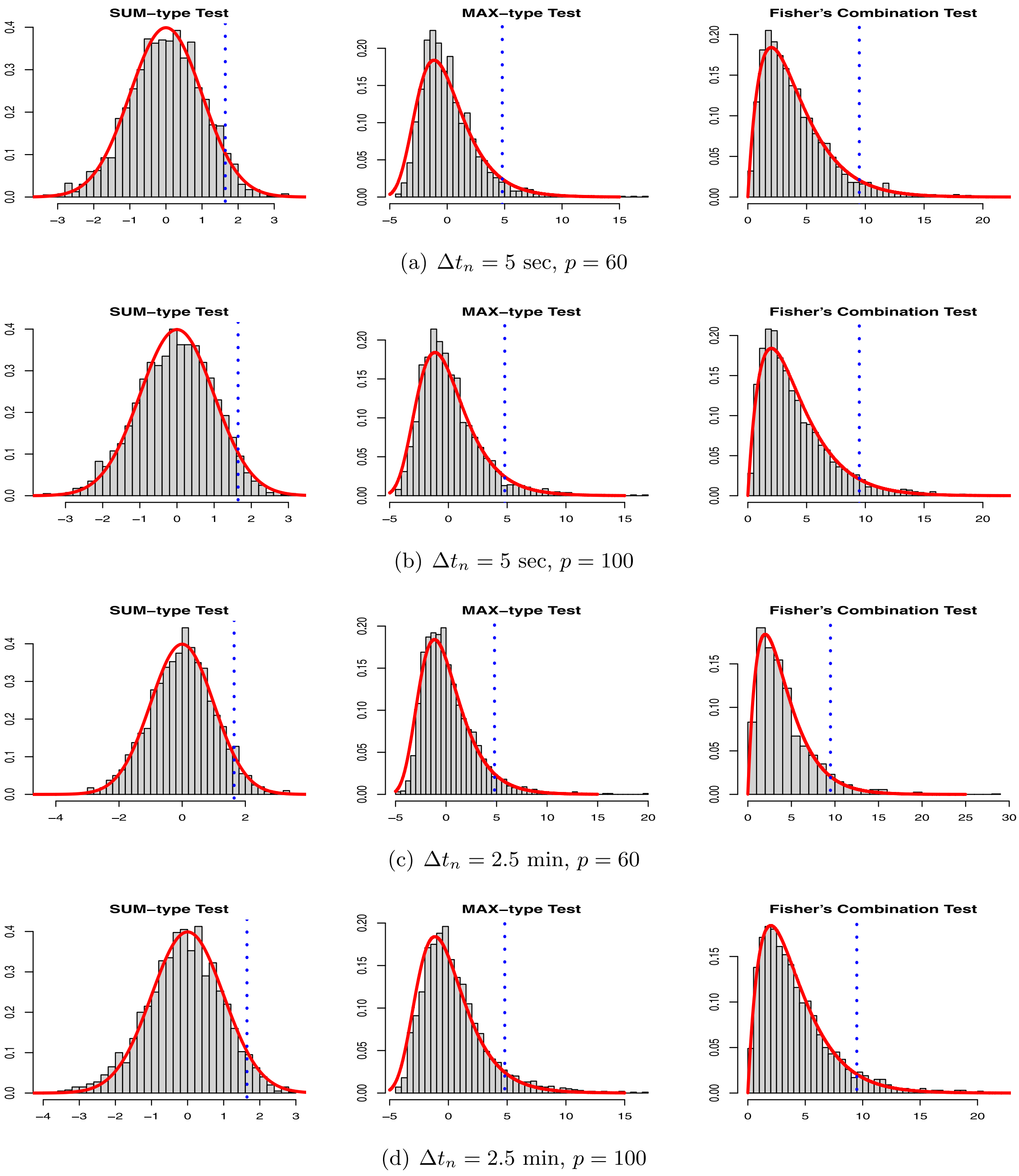
We explore three different testing scenarios, i.e., S1, S2 and S3, which correspond to dense alternative, sparse alternative, and null hypothesis, respectively. In this way, the performance of our methodologies can be comprehensively examined. The detailed specifications of S1, S2 and S3 are described as follows.

(S1) Dense alternative: Y_t is OEF; group a variables are the stocks which do not belong to any sectors, i.e.,

$$X_{a,t} = \left(X_t^{\text{SP100-1}}, X_t^{\text{SP100-2}}, X_t^{\text{SP100-3}}, X_t^{\text{SP100-4}}, X_t^{\text{SP100-5}}\right)^{\intercal};$$

group b variables are the rest of the 80 selected components of S&P 100 Index, i.e.,

$$X_{b,t} = \left(X_t^{\text{SP100-6}}, X_t^{\text{SP100-7}}, \dots, X_t^{\text{SP100-80}}\right)^{\intercal}.$$



(a) $\Delta t_n = 5$ sec, $p = 60$

(b) $\Delta t_n = 5$ sec, $p = 100$

(c) $\Delta t_n = 2.5$ min, $p = 60$

(d) $\Delta t_n = 2.5$ min, $p = 100$

Fig. 4.1. Limiting null distributions of the three proposed tests. Note: This figure shows the limiting null distributions of the three proposed tests, including T_{SUM}, T_{MAX} and T_{FC} which are defined in (3.11), (3.23) and (3.24), respectively. The red solid curves denote the benchmark of the limiting null distribution. The blue dashed lines denote the critical values of the tests.

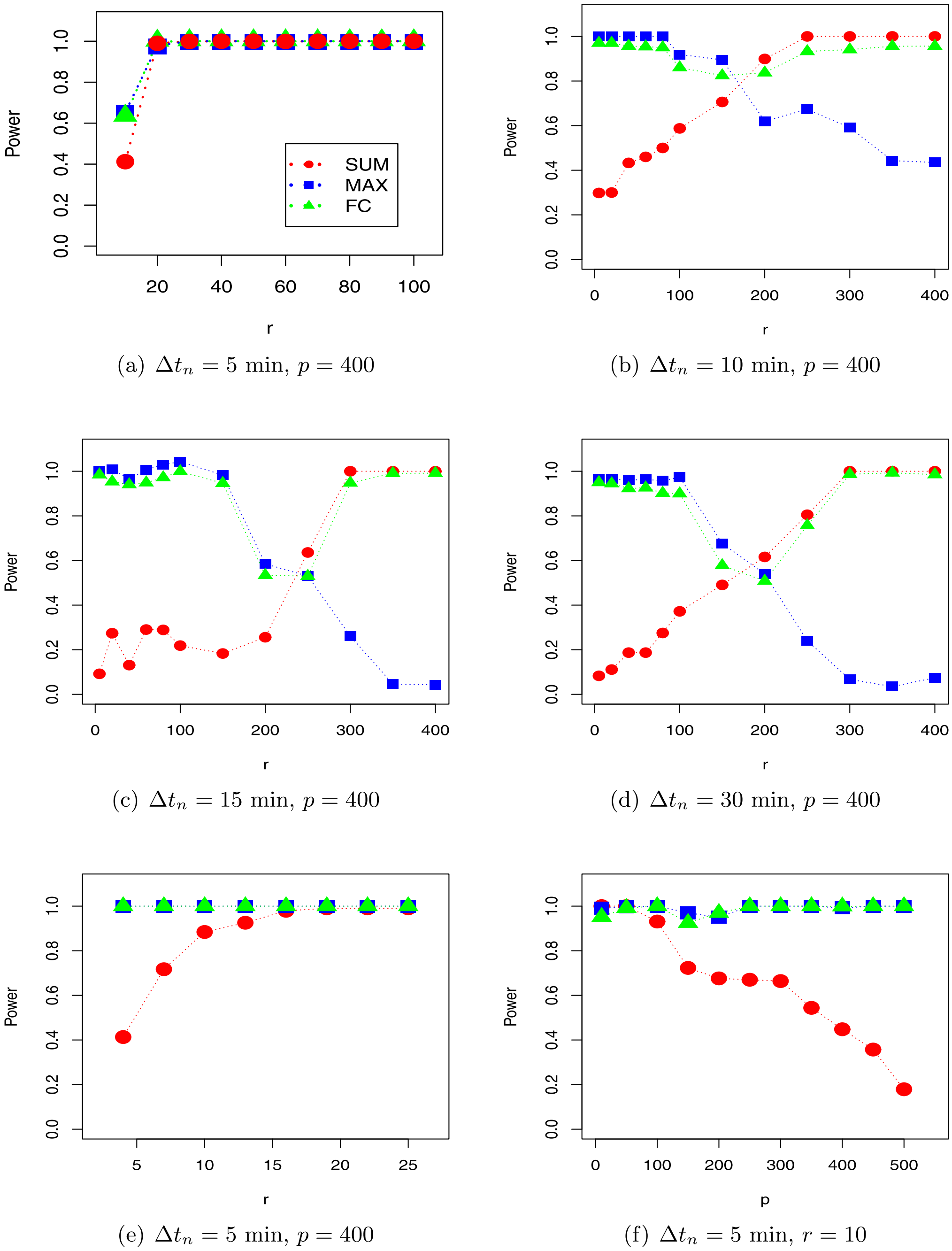
In this scenario, $q = 5$ and $p = 75$. Given the construction of OEF, one can expect most of the regression coefficients associated with $X_{b,t}$ are significantly differently from zero. Hence, this is a "Dense" alternative.

(S2) Sparse alternative: Y_t is XLU; group a variables are the stocks which do not belong to any sectors, i.e.,

$$X_{a,t} = \left(X_t^{\text{SP100-1}}, X_t^{\text{SP100-2}}, X_t^{\text{ SP100-3}}, X_t^{\text{SP100-4}}, X_t^{\text{SP100-5}}\right)^{\intercal};$$

group b variables are the rest of the 80 selected components of S&P 100 Index, i.e.,

$$X_{b,t} = \left(X_t^{\text{SP100-6}}, X_t^{\text{SP100-7}}, \ldots, X_t^{\text{SP100-80}}\right)^{\intercal}.$$



(a) $\Delta t_n = 5$ min, $p = 400$

(b) $\Delta t_n = 10$ min, $p = 400$

(c) $\Delta t_n = 15$ min, $p = 400$

(d) $\Delta t_n = 30$ min, $p = 400$

(e) $\Delta t_n = 5$ min, $p = 400$

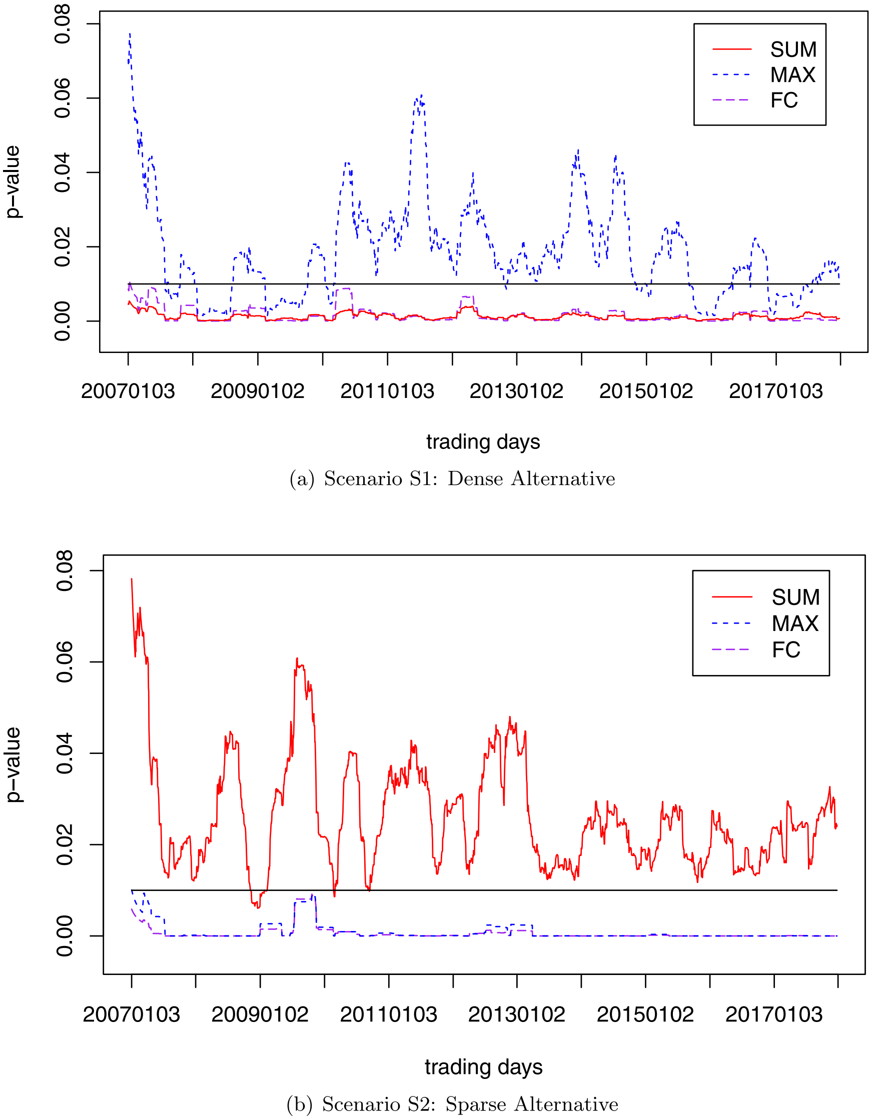
(f) $\Delta t_n = 5$ min, $r = 10$

Fig. 4.2. Empirical power of the three proposed tests. Note: This figure shows the empirical power of the three proposed tests, including T_{SUM}, T_{MAX} and T_{FC} which are defined in (3.11), (3.23) and (3.24), respectively. In the figure, the red circle and "SUM" denote the sum-type test T_{SUM}, the blue square and "MAX" denote the max-type test T_{MAX}, the green triangle and "FC" denote the Fisher's combination test T_{FC}.

In this scenario, $q = 5$ and $p = 75$. Under the construction of XLU ETF, the three regression coefficients associated with the utility stocks ($X_t^{\text{SP100-6}}$, $X_t^{\text{SP100-7}}$, and $X_t^{\text{SP100-8}}$) should be none-zero while the remaining majority $\beta^{(k)}$ with group b covariates are close to zero. Thus this case is called "Sparse" alternative.

(S3) Null hypothesis where all $\beta^{(k)}$'s associated with the group b variables are zero: Y_t is XLU; group a variables are the stocks which do not belong to any sectors plus the 3 stocks belonging to the Sector of Utility, i.e.,

$$X_{a,t} = \left(X_t^{\text{SP100-1}}, X_t^{\text{SP100-2}}, X_t^{\text{ SP100-3}}, X_t^{\text{SP100-4}}, X_t^{\text{SP100-5}}, X_t^{\text{SP100-6}}, X_t^{\text{SP100-7}}, X_t^{\text{SP100-8}}\right)^{\intercal};$$



(a) Scenario S1: Dense Alternative

(b) Scenario S2: Sparse Alternative

Fig. 5.1. p-values of the three proposed tests with real data (Alternative hypothesis). Note: This figure shows the time series of the p-values of the three proposed tests (i.e., sum-type test, max-type test and Fisher's combination test) under Scenario S1 and Scenario S2, respectively.

group b variables are the rest of the 80 selected components of S&P 100 Index, i.e.,

$$X_{b,t} = \left(X_t^{\text{SP100-9}}, X_t^{\text{SP100-10}}, \ldots, X_t^{\text{SP100-80}}\right)^\intercal .$$

In this scenario, $q = 8$ and $p = 72$. When Y is an utility-sector ETF, one expects the regression coefficients associated with non-utility stocks are all close to zero. This scenario resembles the null hypothesis of all $\beta^{(k)}$'s associated with the group b variables are zero.

The time series of the p-values of the three proposed tests are shown in Figs. 5.1-(a), -(b), and 5.2 for Scenarios S1, S2 and S3, respectively.

At the significance level of 0.01, Fig. 5.1-(a) shows that when most of the regression coefficients in group b covariates are none-zero (aka, dense), the sum-type test and the Fisher's combination test correctly rejected the null hypothesis whereas the max-type test failed. Fig. 5.1-(b) shows that the sum-type test failed to detect the small number of none-zero regression coefficients, whereas the max-type test and the Fisher's combination test perform well. This is expected, as the sum-type test is not suitable for sparse alternatives. Both Fig. 5.1(a)–(b) show that, our Fisher's combination test is robust to both sparse and dense alternatives in the real data analysis.

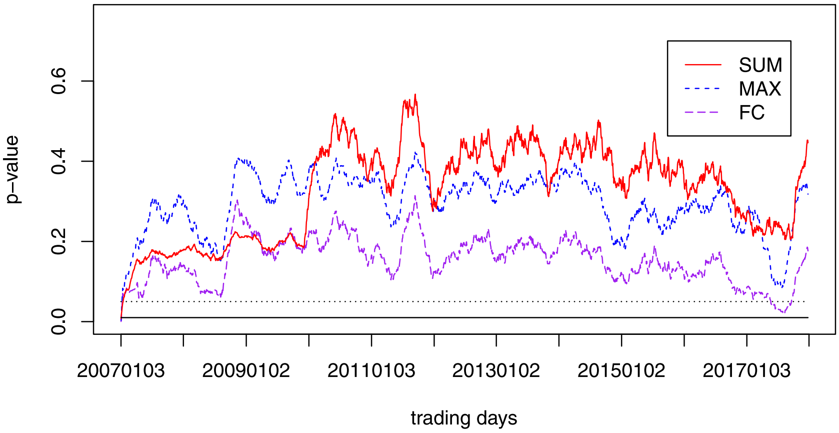


Fig. 5.2. p-values of The three proposed tests with real data (Null hypothesis). Note: This figure shows the time series of the p-values of the three proposed tests (i.e., sum-type test, max-type test and Fisher's combination test) under Scenario S3, which corresponds to the null hypothesis.

Fig. 5.2 further illustrates that all three proposed tests did not have sufficient evidence to reject the null hypothesis. In other words, all three tests correctly "detected" group b regression coefficients being almost all close to zero. This finding aligns with our prediction, as we intentionally set S3 as the null hypothesis to showcase the effectiveness of the proposed tests.

5.2. Example of high frequency factor model

There are a large number of high frequency factors proposed in the field of financial econometrics, see, e.g., Aït-Sahalia et al. (2020) and Aleti (2022). A natural question then arises: can these newly generated factors contain more information about the test assets? In other words, is it possible that the traditional factors (as in Fama–French three-factor model, Fama–French five-factor model, etc.) can achieve the same performance as the combination of the 200+ new factors? Is the incremental impact of high frequency (HF) factors time-varying? If so, can we detect the time periods during which the HF factors add more signal to the traditional Fama–French factor model?

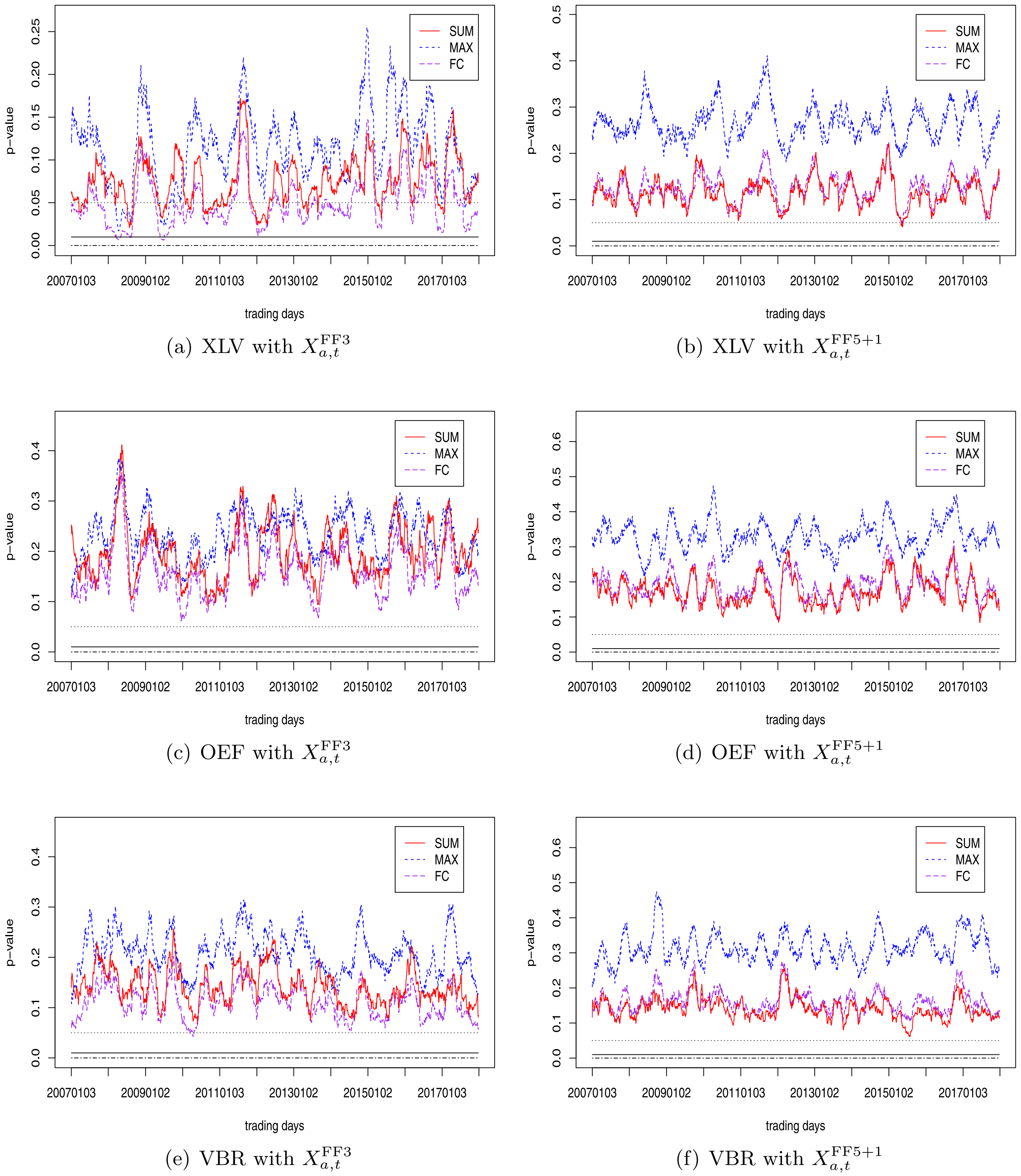
Following our discussion in Remark 5 about signal strength, we consider the following model designs:

1. Response variable Y_t is the test asset;
2. Group a variable $X_{a,t}$ is the set of factors in traditional factor model (FF3, FF5, FF5+1);
3. Group b variable $X_{b,t}$ is the newly generated high-frequency (HF) factors of which the number is large.

We test the impact of $X_{b,t}$ (HF factors) on the test asset Y_t given that $X_{a,t}$ (traditional Fama–French-type factors) are already in the model. A small p-value (below the significance level $\alpha = 0.05$) indicates that the group b variables may contribute additional variations in the test asset compared to the traditional factor model. Conversely, a larger p-value of this test (i.e. $p > 0.05$) suggests that the traditional factor model is as effective as the newly generated factors in explaining the test asset's variations.

In this study, we choose the following ETFs as the test assets: (i) iShares S&P 100 ETF (OEF), which tracks the performance of the 100 large-cap stocks, (ii) Vanguard Size/Style ETFs, including Small-Cap Value ETF (VBR), Small-Cap Growth ETF (VBK), Large-Cap Value ETF (VTV), Large-Cap Growth ETF (VUG); (iii) Sector SDPR ETFs, including Materials (XLB), Energy (XLE), Financial (XLF), Industrials (XLI), Information Technology (XLK), Consumer Staples (XLP), Utilities (XLU), Health Care (XLV) and Consumer Discretionary (XLY). The intraday asset prices of these ETFs are downloaded from the TAQ database of NYSE. For the group a variables, we consider three cases based on the Fama–French $5 + 1$ factors $(X_t^{\mathrm{MKT}}, X_t^{\mathrm{SMB}}, X_t^{\mathrm{HML}}, X_t^{\mathrm{RMW}}, X_t^{\mathrm{CMA}}, X_t^{\mathrm{UMD}})$, of which the intraday returns are constructed and shared by the author of Aleti (2022): (i) FF3: $X_{a,t}^{\mathrm{FF3}} = (X_t^{\mathrm{MKT}}, X_t^{\mathrm{SMB}}, X_t^{\mathrm{HML}})$; (ii) FF5: $X_{a,t}^{\mathrm{FF5}} = (X_t^{\mathrm{MKT}}, X_t^{\mathrm{SMB}}, X_t^{\mathrm{HML}}, X_t^{\mathrm{RMW}}, X_t^{\mathrm{CMA}})$; (iii) FF5+1: $X_{a,t}^{\mathrm{FF5+1}} = (X_t^{\mathrm{MKT}}, X_t^{\mathrm{SMB}}, X_t^{\mathrm{HML}}, X_t^{\mathrm{RMW}}, X_t^{\mathrm{CMA}}, X_t^{\mathrm{UMD}})$. For the group b variables, we choose the 272 high frequency factors, which are also constructed and shared by the author of Aleti (2022). Similar to the first example, $\Delta t_n = 15$ min.

We first examine the p-values over the course of 11 years from 2007 to 2017. Setting $\mathcal{T} = 11$ years, Table 5.1 shows that the HF factors add significant information ($\alpha = 0.05$) when explaining the variation of any test asset in our consideration. This is demonstrated from the extremely small p-values in the Fisher's combination test, regardless of the $X_{a,t}$ variables being the Fame-French 3 factors (FF3), Fama–French 5 factors (FF5), or the FF 5 factors plus the momentum factor (FF5+1). The Sum-type and Max-type tests also agree on almost all cases, with the exception of the small-cap value fund ETF (VBR) as the response variable. When pinning down to VBR though, one finds that the SUM-type test and the MAX-type test differ when $X_{a,t}$ =FF5+1 factors, with the former test rejecting the null hypothesis ($H_0 : \beta_{b,t} = 0$) whereas the latter not rejecting. This suggests that after accounting for the impact of the FF5+1 factors on VBR, the 272 HF factors resemble the dense alternative, i.e. many small-impact factors.



(a) XLV with $X_{a,t}^{\mathrm{FF3}}$

(b) XLV with $X_{a,t}^{\mathrm{FF5+1}}$

(c) OEF with $X_{a,t}^{\mathrm{FF3}}$

(d) OEF with $X_{a,t}^{\mathrm{FF5+1}}$

(e) VBR with $X_{a,t}^{\mathrm{FF3}}$

(f) VBR with $X_{a,t}^{\mathrm{FF5+1}}$

Fig. 5.3. Time series of *p*-values when HF factors are added to the traditional Fama–French factor models. Notes. The black solid line denotes 0.01; the black dotted line above the black solid line denotes 0.05; the black dashed line underneath the black solid line denotes 0.

On the monthly level ($\mathcal{T}$ = 1 month), Fig. 5.3 shows the time varying nature in p-values. For the health care fund (XLV), after all FF 5 factors plus momentum factor are incorporated into the model, none of the 272 HF factors add significant signal for the variation in XLV returns. However, with only FF 3 factors in the model, many low-impact HF factors plus a few high-impact factors all provide incremental influence to XLV in mid-year 2009 (p-values < 0.05 in both MAX- and SUM-type test); in other months, say, the first quarter of 2012, the estimated p-values are less than 0.05 in the SUM-type test but greater than 0.05 in the MAX-type test, revealing the presence of many low-impact HF factors (while lacking high-impact HF factors) influencing XLV.

For the S&P100 tracking fund OEF and the small value fund (VBR) in Fig. 5.3, almost all the p-values are greater than 0.05, suggesting that the original Fama–French 3 factors are sufficient to explain the variation in OEF and VBR on a monthly basis, with

Table 5.1
p-values when HF factors are added to the traditional Fama–French factor models.

Test asset	SUM-type test			MAX-type test			Fisher's combination test		
	$X_{a,t}^{\mathrm{FF3}}$	$X_{a,t}^{\mathrm{FF5}}$	$X_{a,t}^{\mathrm{FF5+1}}$	$X_{a,t}^{\mathrm{FF3}}$	$X_{a,t}^{\mathrm{FF5}}$	$X_{a,t}^{\mathrm{FF5+1}}$	$X_{a,t}^{\mathrm{FF3}}$	$X_{a,t}^{\mathrm{FF5}}$	$X_{a,t}^{\mathrm{FF5+1}}$
OEF	2.6×10^{-14}	3.3×10^{-12}	6.7×10^{-10}	3.3×10^{-16}	4.9×10^{-11}	1.1×10^{-8}	$< 10^{-16}$	$<10^{-16}$	3.3×10^{-16}
VBR	0.011	0.022	0.036	0.003	0.048	0.076	3.8×10^{-4}	0.008	0.019
VBK	0.004	0.020	0.016	2.3×10^{-5}	1.9×10^{-4}	0.001	1.7×10^{-6}	5.3×10^{-5}	1.9×10^{-4}
VTV	1.8×10^{-11}	3.0×10^{-10}	2.4×10^{-9}	7.6×10^{-11}	5.2×10^{-10}	1.6×10^{-7}	$<10^{-16}$	$<10^{-16}$	1.4×10^{-14}
VUG	$<10^{-16}$	$<10^{-16}$	3.3×10^{-16}	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$< 10^{-16}$	$<10^{-16}$
XLB	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$
XLE	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$
XLF	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$
XLI	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$
XLK	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$
XLP	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$
XLU	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$
XLV	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$
XLY	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$

Notes. This table reports the *p*-values of the proposed tests for $X_{b,t}$ under different combinations of test assets and traditional factor models over the full time period of $\mathcal{T} = 11$ years (from Jan 2007 to Dec 2017).

the 272 HF factors adding little contribution. The only exception is around the May 2010 (flash crash), during which the Fisher's combination test detected some signal from the HF factors. We notice that in all three panels with X_a = FF5+1, the monthly (insignificant) *p*-value series under SUM test and Fisher's combination test are close to each other. In other words, with all Fama–French 5 factors and momentum factor included in the model, HF factors are more like many tiny and insignificant signals on a monthly level.

We should mention that the statistically insignificance on a monthly scale (as in Fig. 5.3) does not contradict the highly significant influence of HF factors over the span of 11 years (as in Table 5.1). Under H_0, the monthly p-values should be uniformly distributed over (0,1). The fact that our monthly p-values in Fig. 5.3 hover around 0.1–0.2 is indicative that the *p*-value is highly significant when being inspected over the course of 11 years. In other words, if we conduct a Fisher's combination test on the 132 (11 years *x* 12 months/year) monthly p-values, we would see similar results as in Table 5.1.

We also caution that our results are about the statistical significance of the HF factors, not their economics significance. For example, even if the HF factors are statistically insignificant in the presence of FF5+1 factors, HF factors may still add economic profit in a trading strategy. The practical value of our proposed tests is to provide a statistical guidance to separating high-impact, low-impact, and no-impact factors/signals. It also help identify when HF factors "move" in and out of the factor models.

6. Conclusion

In this paper, we developed three tests for the regression coefficient processes in the high dimensional and high frequency regression, including the sum-type test, max-type test, and the Fisher's combination test. The limiting null distributions of these three proposed tests are derived and the asymptotic behavior of their powers are also analyzed. The max-type test can work well with the sparse alternative, where there are few non-zero elements in the high-dimensional beta processes. In contrast, the sum-type test can work well with the dense alternative. As the combination of the sum-type test and max-type test, Fisher's combination test is robust to both sparse and dense alternatives. When applied to high dimensional high frequency setting, these three tests help us identify additional HF signals for test assets. The tests also permit separation of high-signal versus (many) low-signal factors, and provide guidance to locate when the high-frequency factors "move" in and out of the usual factor models.

Acknowledgments

The authors thank Carsten Chong, Chenxu Li, Yingying Li, Cheng Liu, Jun Liu, Liangjun Su, Shouyang Wang, Xinghua Zheng, Yong Zhou, the editor, the associate editor and three anonymous referees for helpful comments and discussions, as well as seminar and conference participants at East China Normal University, Hong Kong University of Science and Technology, Nanjing Audit University, Peking University, Shanghai University of Finance and Economics, Tsinghua University, and the 12th Chinese Conference of Probability and Statistics, SoFiE 2023 Annual Conference, Workshop on Theory and Application of Time-Varying Econometrics Models. Chen is supported by School of Economics, Singapore Management University, and National Natural Science Foundation of China (Grant 12101335). Feng is partially supported by Shenzhen Wukong Investment Company, Tianjin Science Fund, China for Outstanding Young Scholar (23JCJQJC00150), the Fundamental Research Funds for the Central Universities, China under Grant No. ZB22000105 and 63233075, the China National Key R&D Program (Grant Nos. 2019YFC1908502, 2022YFA1003703, 2022YFA1003802, 2022YFA1003803) and the National Natural Science Foundation of China Grants (Nos. 12271271, 11925106, 12231011, 11931001 and 11971247). The authors would also like to thank the United States National Science Foundation under Grants DMS-2015544 and DMS-2413952 (Mykland), and DMS-2015530 and DMS-2413953 (Zhang).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2024.105812.

References

Aït-Sahalia, Y., Fan, J., Xiu, D., 2010. High-frequency covariance estimates with noisy and asynchronous financial data. J. Amer. Statist. Assoc. 105 (492), 1504–1517.

Aït-Sahalia, Y., Kalnina, I., Xiu, D., 2020. High-frequency factor models and regressions. J. Econometrics 216 (1), 86–105.

Aït-Sahalia, Y., Kimmel, R., 2007. Maximum likelihood estimation of stochastic volatility models. J. Financ. Econ. 83, 413–452.

Aït-Sahalia, Y., Xiu, D., 2017. Using principal component analysis to estimate a high dimensional factor model with high-frequency data. J. Econometrics 201 (2), 384–399.

Aït-Sahalia, Y., Xiu, D., 2019. Principal component analysis of high-frequency data. J. Amer. Statist. Assoc. 114 (525), 287–303.

Aleti, S., 2022. The high-frequency factor zoo. Available at SSRN 4021620.

Andersen, T.G., Bollerslev, T., Diebold, F.X., Wu, J., 2005. A framework for exploring the macroeconomic determinants of systematic risk. Amer. Econ. Rev. 95 (2), 398–404.

Bai, Z., Saranadasa, H., 1996. Effect of high dimension: by an example of a two sample problem. Statist. Sinica 311–329.

Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A., Shephard, N., 2011. Multivariate realised kernels: consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. J. Econometrics 162 (2), 149–169.

Barndorff-Nielsen, O.E., Shephard, N., 2004. Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics. Econometrica 72 (3), 885–925.

Bollerslev, T., Meddahi, N., Nyawa, S., 2019. High-dimensional multivariate realized volatility estimation. J. Econometrics 212 (1), 116–136.

Cai, T.T., Liu, W., Xia, Y., 2014. Two-sample test of high dimensional means under dependence. J. R. Stat. Soc. Ser. B Stat. Methodol. 76 (2), 349–372.

Chen, D., Feng, L., 2022. Asymptotic independence of the quadratic form and maximum of independent random variables with applications to high-dimensional tests. arXiv preprint arXiv:2204.08628.

Chen, D., Mykland, P.A., Zhang, L., 2020. The five trolls under the bridge: Principal component analysis with asynchronous and noisy high frequency data. J. Amer. Statist. Assoc. 115 (532), 1960–1977.

Chen, D., Mykland, P.A., Zhang, L., 2024. Realized regression with asynchronous and noisy high frequency and high dimensional data. J. Econometr. 239 (2), 105446.

Chen, S.X., Qin, Y.-L., 2010. A two-sample test for high-dimensional data with applications to gene-set testing. Ann. Statist. 38 (2), 808–835.

Dai, C., Lu, K., Xiu, D., 2019. Knowing factors or factor loadings, or neither? Evaluating estimators of large covariance matrices with noisy and asynchronous data. J. Econometrics 208 (1), 43–79.

Fama, E.F., French, K.R., 2017. International tests of a five-factor asset pricing model. J. Financ. Econom. 123 (3), 441–463.

Fan, J., Furger, A., Xiu, D., 2016. Incorporating global industrial classification standard into portfolio allocation: A simple factor-based large covariance matrix estimator with high-frequency data. J. Bus. Econom. Statist. 34 (4), 489–503.

Fan, J., Kim, D., 2018. Robust high-dimensional volatility matrix estimation for high-frequency factor model. J. Amer. Statist. Assoc. 113 (523), 1268–1283.

Fan, J., Liao, Y., Mincheva, M., 2013. Large covariance estimation by thresholding principal orthogonal complements. J. R. Stat. Soc. Ser. B Stat. Methodol. 75 (4), 603–680.

Fan, J., Liao, Y., Yao, J., 2015. Power enhancement in high-dimensional cross-sectional tests. Econometrica 83 (4), 1497–1541.

Fathi, M., 2021. Higher-order stein kernels for Gaussian approximation. Studia Math. 256, 241–258.

Feng, L., Jiang, T., Li, X., Liu, B., 2022. Asymptotic independence of the sum and maximum of dependent random variables with applications to high-dimensional tests. arXiv:2205.01638.

He, Y., Xu, G., Wu, C., Pan, W., 2021. Asymptotically independent U-statistics in high-dimensional testing. Ann. Statist. 49 (1), 154–181.

Jacod, J., Li, Y., Mykland, P., Podolskij, M., Vetter, M., 2009. Microstructure noise in the continuous case: the pre-averaging approach. Stochastic Process. Appl. 119, 2149–2276.

Jacod, J., Protter, P., 2011. Discretization of Processes, vol. 67, Springer Science & Business Media.

Kim, D., Liu, Y., Wang, Y., 2018. Large volatility matrix estimation with factor-based diffusion model for high-frequency financial data.

Kim, D., Shin, M., 2022. High-dimensional time-varying coefficient estimation. Available at SSRN 4037351.

Kong, X.-B., 2017. On the number of common factors with high-frequency data. Biometrika 104 (2), 397–410.

Kong, X.-B., 2018. On the systematic and idiosyncratic volatility with large panel high-frequency data.

Kong, X.-B., Lin, J.-G., Liu, C., Liu, G.-Y., 2021. Discrepancy between global and local principal component analysis on large-panel high-frequency data. J. Amer. Statist. Assoc. (just-accepted), 1–32.

Kong, X.-B., Liu, C., 2018. Testing against constant factor loading matrix with large panel high-frequency data. J. Econometrics 204 (2), 301–319.

Ledoux, M., Nourdin, I., Peccati, G., 2015. Stein's method, logarithmic Sobolev and transport inequalities. Geom. Funct. Anal. 25 (1), 256–306.

Littell, R.C., Folks, J.L., 1971. Asymptotic optimality of Fisher's method of combining independent tests. J. Amer. Statist. Assoc. 66 (336), 802–806.

Littell, R.C., Folks, J.L., 1973. Asymptotic optimality of Fisher's method of combining independent tests II. J. Amer. Statist. Assoc. 68 (341), 193–194.

Liu, Y., Zhang, S., Ma, S., Zhang, Q., 2020. Tests for regression coefficients in high dimensional partially linear models. Statist. Probab. Lett. 163, 108772.

Mykland, P.A., Zhang, L., 2006. ANOVA for diffusions and Ito processes. Ann. Statist. 34 (4), 1931–1963.

Mykland, P.A., Zhang, L., 2009. Inference for continuous semimartingales observed at high frequency. Econometrica 77 (5), 1403–1445.

Mykland, P.A., Zhang, L., 2012. The econometrics of high frequency data. Statist. Methods Stoch. Differ. Equ. 124, 109.

Mykland, P.A., Zhang, L., 2017. Assessment of uncertainty in high frequency data: The observed asymptotic variance. Econometrica 85 (1), 197–231.

Mykland, P.A., Zhang, L., Chen, D., 2019. The algebra of two scales estimation, and the S-TSRV: high frequency estimation that is robust to sampling times. J. Econometrics 208 (1), 101–119.

Pelger, M., 2019. Large-dimensional factor modeling based on high-frequency observations. J. Econometrics 208 (1), 23–42.

Pelger, M., 2020. Understanding systematic risk: A high-frequency approach. J. Finance 75 (4), 2179–2220.

Reiß, M., Todorov, V., Tauchen, G., 2015. Nonparametric test for a constant beta between Itô semi-martingales based on high-frequency data. Stochastic Process. Appl. 125 (8), 2955–2988.

Shin, M., Kim, D., 2023. Robust high-dimensional time-varying coefficient estimation. arXiv preprint arXiv:2302.13658.

Srivastava, M.S., 2009. A test for the mean vector with fewer observations than the dimension under non-normality. J. Multivariate Anal. 100 (3), 518–532.

Srivastava, M.S., Du, M., 2008. A test for the mean vector with fewer observations than the dimension. J. Multivariate Anal. 99 (3), 386–402.

Srivastava, M.S., Katayama, S., Kano, Y., 2013. A two sample test in high dimensional data. J. Multivariate Anal. 114, 349–358.

Xu, G., Lin, L., Wei, P., Pan, W., 2016. An adaptive two-sample test for high-dimensional means. Biometrika 103 (3), 609–624.

Zhang, L., 2006. Efficient estimation of stochastic volatility using noisy observations: A multi-scale approach. Bernoulli 12 (6), 1019–1043.

Zhang, L., 2011. Estimating covariation: Epps effect, microstructure noise. J. Econometrics 160 (1), 33–47.

Zhang, L., Mykland, P.A., Aït-Sahalia, Y., 2005. A tale of two time scales: determining integrated volatility with noisy high-frequency data. J. Amer. Statist. Assoc. 100, 1394–1411.